

LNC5 15327

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

Pattern Recognition

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

27 Part XXVII



Lecture Notes in Computer Science

15327

Founding Editors


Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

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 XXVII

Editors

Apostolos Antonacopoulos 
University of Salford
Salford, UK

Rama Chellappa 
Johns Hopkins University
Baltimore, MD, USA

Saumik Bhattacharya 
IIT Kharagpur
Kharagpur, India

Subhasis Chaudhuri 
Indian Institute of Technology Bombay
Mumbai, India

Cheng-Lin Liu 
Chinese Academy of Sciences
Beijing, China

Umapada Pal 
Indian Statistical Institute Kolkata
Kolkata, India

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-78397-5

ISBN 978-3-031-78398-2 (eBook)

<https://doi.org/10.1007/978-3-031-78398-2>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

If disposing of this product, please recycle the paper.

President's Address

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

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

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

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

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

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

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

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

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

September 2024

Arjan Kuijper
President of the IAPR

Preface

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

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

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

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

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

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

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

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

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

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

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

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

October 2024

Umapada Pal
Josef Kittler
Anil Jain

Organization

General Chairs

Umapada Pal
Josef Kittler
Anil Jain

Indian Statistical Institute, Kolkata, India
University of Surrey, UK
Michigan State University, USA

Program Chairs

Apostolos Antonacopoulos
Subhasis Chaudhuri
Rama Chellappa
Cheng-Lin Liu

University of Salford, UK
Indian Institute of Technology, Bombay, India
Johns Hopkins University, USA
Institute of Automation, Chinese Academy of
Sciences, China

Publication Chairs

Ananda S. Chowdhury
Wataru Ohyama

Jadavpur University, India
Tokyo Denki University, Japan

Competition Chairs

Richard Zanibbi
Lianwen Jin
Laurence Likforman-Sulem

Rochester Institute of Technology, USA
South China University of Technology, China
Télécom Paris, France

Workshop Chairs

P. Shivakumara
Stephanie Schuckers
Jean-Marc Ogier
Prabir Bhattacharya

University of Salford, UK
Clarkson University, USA
Université de la Rochelle, France
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 Lladós	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. Ambasmudram	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	Jiixin Zhang
Yuliang Liu	Yiwu Zhong
Yongxin Shi	

Reviewers (Conference Papers)

Aakanksha Aakanksha
 Aayush Singla
 Abdul Muqet
 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
 Anwasha Sengupta
 Archisman Adhikary
 Arjan Kuijper
 Arnab Kumar Das

Arnav Bhavsar	Bin-Bin Jia
Arnav Varma	Binbin Yong
Arpita Dutta	Bindita Chaudhuri
Arshad Jamal	Bindu Madhavi Tummala
Artur Jordao	Binh M. Le
Arunkumar Chinnaswamy	Bi-Ru Dai
Aryan Jadon	Bo Huang
Aryaz Baradarani	Bo Jiang
Ashima Anand	Bob Zhang
Ashis Dhara	Bowen Liu
Ashish Phophalia	Bowen Zhang
Ashok K. Bhateja	Boyang Zhang
Ashutosh Vaish	Boyu Diao
Ashwani Kumar	Boyun Li
Asifuzzaman Lasker	Brian M. Sadler
Atefeh Khoshkhahtinat	Bruce A. Maxwell
Athira Nambiar	Bryan Bo Cao
Attilio Fiandrotti	Buddhika L. Semage
Avandra S. Hemachandra	Bushra Jalil
Avik Hati	Byeong-Seok Shin
Avinash Sharma	Byung-Gyu Kim
B. H. Shekar	Caihua Liu
B. Uma Shankar	Cairong Zhao
Bala Krishna Thunakala	Camille Kurtz
Balaji Tk	Carlos A. Caetano
Balázs Pálffy	Carlos D. Martá-Nez-Hinarejos
Banafsheh Adami	Ce Wang
Bang-Dang Pham	Cevahir Cigla
Baochang Zhang	Chakravarthy Bhagvati
Baodi Liu	Chandrakanth Vipparla
Bashirul Azam Biswas	Changchun Zhang
Beiduo Chen	Changde Du
Benedikt Kottler	Changkun Ye
Beomseok Oh	Changxu Cheng
Berkay Aydin	Chao Fan
Berlin S. Shaheema	Chao Guo
Bertrand Kerautret	Chao Qu
Bettina Finzel	Chao Wen
Bhavana Singh	Chayan Halder
Bibhas C. Dhara	Che-Jui Chang
Bilge Günsel	Chen Feng
Bin Chen	Chenan Wang
Bin Li	Cheng Yu
Bin Liu	Chenghao Qian
Bin Yao	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 Coquenat
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
Dipanjoyoti 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 Binamakhshen
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 Otroschi 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
Jiyang 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
Liviú-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 Kämpel
Martina Pastorino
Marwan Turki
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	Palash Ghosal
Narayan Vetrekar	Pallav Dutta
Narendra D. Londhe	Paolo Rota
Nathalie Girard	Paramanand Chandramouli
Nati Ofir	Paria Mehrani
Naval Kishore Mehta	Parth Agrawal
Nazmul Shahadat	Partha Basuchowdhuri
Neeti Narayan	Patrick Horain
Neha Bhargava	Pavan Kumar
Nemanja Djuric	Pavan Kumar Anasosalu Vasu
Newlin Shebiah R.	Pedro Castro
Ngo Ba Hung	Peipei Li
Nhat-Tan Bui	Peipei Yang
Niaz Ahmad	Peisong Shen
Nick Theisen	Peiyu Li
Nicolas Passat	Peng Li
Nicolas Ragot	Pengfei He
Nicolas Sidere	Pengrui Quan
Nikolaos Mitianoudis	Pengxin Zeng
Nikolas Ebert	Pengyu Yan
Nilah Ravi Nair	Peter Eisert
Nilesh A. Ahuja	Petra Gomez-Krämer
Nilkanta Sahu	Pierrick Bruneau
Nils Murrugarra-Llerena	Ping Cao
Nina S. T. Hirata	Pingping Zhang
Ninad Aithal	Pintu Kumar
Ning Xu	Pooja Kumari
Ningzhi Wang	Pooja Sahani
Niraj Kumar	Prabhu Prasad Dev
Nirmal S. Punjabi	Pradeep Kumar
Nisha Varghese	Pradeep Singh
Norio Tagawa	Pranjal Sahu
Obaidullah Md Sk	Prasun Roy
Oguzhan Ulucan	Prateek Keserwani
Olfa Mechi	Prateek Mittal
Oliver Tüselmann	Praveen Kumar Chandaliya
Orazio Pontorno	Praveen Tirupattur
Oriol Ramos Terrades	Pravin Nair
Osman Akin	Preeti Gopal
Ouadi Beya	Preety Singh
Ozge Mercanoglu Sincan	Prem Shanker Yadav
Pabitra Mitra	Prerana Mukherjee
Padmanabha Reddy Y. C. A.	Prerna A. Mishra
Palaash Agrawal	Prianka Dey
Palaiahnakote Shivakumara	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
Ruwei 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
Siladitya 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 Napol on
 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 B chner

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

Time-Series Representation Learning via Heterogeneous Spatial-Temporal Contrasting for Remaining Useful Life Prediction	1
<i>Zhixin Huang, Yujiang He, Chandana Priya Nivarthi, Bernhard Sick, and Christian Gruhl</i>	
CCPL: Cross-Modal Contrastive Protein Learning	22
<i>Jiangbin Zheng and Stan Z. Li</i>	
Box2Flow: Instance-Based Action Flow Graphs from Videos	39
<i>Jiatong Li, Kalliopi Basioti, and Vladimir Pavlovic</i>	
Learning Geometry of Pose Image Manifolds in Latent Spaces Using Geometry-Preserving GANs	56
<i>Shenyuan Liang, Benjamin Beaudett, Pavan Turaga, Saket Anand, and Anuj Srivastava</i>	
MOMA: Contrastive Learning Distills Better Masked Autoencoders	73
<i>Yuchong Yao, Nandakishor Desai, and Marimuthu Palaniswami</i>	
Leveraging Cross-Augmentation Consensus and Conflict for Semi-supervised Semantic Segmentation	89
<i>Junhao Cao, Junyi Chen, Sibor Huang, and Dongyu Zhang</i>	
Event-Aware Multi-component (EMI) Loss for Fraud Detection	105
<i>Tarun Somavarapu, Anand Vir Singh, Maneet Singh, Shraddha Pandey, Shantanu Verma, and Kushagra Agarwal</i>	
A Simple Heuristic for Controlling Human Workload in Learning to Defer	120
<i>Andrew Ponomarev</i>	
Explore Statistical Properties of Undirected Unweighted Networks from Ensemble Models	131
<i>Xunda Zhao, Xing Wu, and Jianjia Wang</i>	
Evaluation of Machine Learning Techniques for Classification of Surface Roughness of Machined Samples using Laser Speckle Imaging Technique	146
<i>Shanta Hardas Patil</i>	

Model Selection with a Shapelet-Based Distance Measure for Multi-source Transfer Learning in Time Series Classification	160
<i>Jiseok Lee and Brian Kenji Iwana</i>	
DACOA: Diffusion-Aligned Coherent Augmentation and Consistency Constraint Strategies for Federated Domain Generalization	176
<i>Guangshuo Wang, Yuesheng Zhu, and Guibo Luo</i>	
Collaborative Domain Alignment for Multi-source Domain Adaptation	192
<i>Yuan Yuan Xu, Meina Kan, Zhilong Ji, Jinfeng Bai, Shiguang Shan, and Xilin Chen</i>	
Edge-Guided and Cross-Scale Feature Fusion Network for Efficient Multi-contrast MRI Super-Resolution	208
<i>Zhiyuan Yang, Bo Zhang, Zhiqiang Zeng, and Si Yong Yeo</i>	
Dual-ResShift: Dual-Input Separated Features Residual Shift Diffusion Model for CTA Image Super-Resolution	219
<i>Feng Jiang, Jing Wen, and Yi Wang</i>	
Multi-Block U-Net for Wind Noise Reduction in Hearing Aids	234
<i>Arth J. Shah, Manish Suthar, and Hemant A. Patil</i>	
A Cascading Approach with Vision Transformers for Age-Related Macular Degeneration Diagnosis and Explainability	250
<i>Ainhoa Osa-Sanchez, Hossam Magdy Balaha, Mahmoud Ali, Mostafa Abdelrahim, Mohmaed Khudri, Begonya Garcia-Zapirain, and Ayman El-Baz</i>	
DCRUNet++: A Depthwise Convolutional Residual UNet++ Model for Brain Tumor Segmentation	266
<i>Yash Sonawane, Maheshkumar H. Kolekar, Agnesh Chandra Yadav, Gargi Kadam, Sanika Tiwarekar, and Dhananjay R. Kalbande</i>	
DrowzEE-G-Mamba: Leveraging EEG and State Space Models for Driver Drowsiness Detection	281
<i>Gourav Siddhad, Sayantan Dey, and Partha Pratim Roy</i>	
EEG-Based Reaction Time Prediction Using Covariance Augmented 2D Convolutional Neural Network	296
<i>Adarsh V. Parekkattil, Sanjeev Kumar Varun, and Tharun Kumar Reddy Bollu</i>	
Uncertainty-RIFA-Net: Uncertainty Aware Robust Information Fusion Attention Network for Brain Tumors Classification in MRI Images	311
<i>Joy Dhar, Kapil Rana, and Puneet Goyal</i>	

Auxiliary Information Guided Segmentation for the Clinical Target
 Volume of Cervical Cancer 328
Shiyun Wang and Yongchao Xu

Synthetic Images with Dense Annotations and Ensemble Learning
 for DFU Segmentation 344
*Pin Xu, Xiongjiang Xiao, Weimin Yuen, Yanyi Li, Kuan Li,
 and Jianping Yin*

SWJEPa: Improving Prostate Cancer Lesion Detection with Shear Wave
 Elastography and Joint Embedding Predictive Architectures 359
*Markus Bauer, Adam Gurwin, Christoph Augenstein,
 Bogdan Franczyk, and Bartosz Malkiewicz*

AdaSVaT: Adaptive Singular Value Thresholding for Adversarial
 Detection in Fundus Images 376
Nirmal Joseph, Sudhish N. George, P. M. Ameer, and Kiran Raja

A New AI System for Precise Grading of HCC Based on Analyzing
 DW-MRI Radiomics and Alpha-fetoprotein as Liver Cancer Clinical
 Marker 392
*Abdelrhman Elkhoully, Ahmed Alksas, Gehad A. Saleh,
 Mohamed Shehata, Abdelrahman Karawia, Mohammed Ghazal,
 Sohail Contractor, and Ayman El-Baz*

Detection of Extremely Sparse Key Instances in Whole Slide Cytology
 Images via Self-supervised One-class Representation Learning 408
*Swarnadip Chatterjee, Orcun Göksel, Nataša Sladoje,
 and Joakim Lindblad*

Hybrid CNN-LSTM Framework for Enhanced Congestive Heart Failure
 Diagnosis: Integrating QRS Detection 422
*Aditya Oza, Sanskriti Patel, Bhavesh Gyanchandani, Abhinav Roy,
 and Santosh Kumar*

A Multimodal MRI-based Framework for Thyroid Cancer Diagnosis
 Using eXplainable Machine Learning 438
*Ahmed Sharafeldeen, Hossam Magdy Balaha, Ali Mahmoud,
 Reem Khaled, Saher Taman, Manar Mansour Hussein,
 Mohammed Ghazal, and Ayman El-Baz*

Author Index 453



Time-Series Representation Learning via Heterogeneous Spatial-Temporal Contrasting for Remaining Useful Life Prediction

Zhixin Huang^(✉), Yujiang He, Chandana Priya Nivarthi, Bernhard Sick,
and Christian Gruhl

Intelligent Embedded Systems, University of Kassel, Kassel, Germany
{zhixin.huang,yujiang.he,chandana.nivarthi,bsick,cgruhl}@uni-kassel.de

Abstract. Classical contrastive learning paradigms rely on manual augmentations like cropping, masking, dropping, or adding noise randomly to create divergent sample views from original data. However, the choice of which method to manipulate samples is often subjective and may destroy the latent pattern of the sample. In response, this paper introduces a novel contrastive learning paradigm without choosing sample view augmentation methods, termed Heterogeneous Spatial-Temporal Representation Contrasting (HSTRC). Instead of sample view augmentation, we employ dual branches with a heterogeneous spatial-temporal flipped structure to extract two distinct hidden feature views from the same source data, which avoids disturbing the original time series. Leveraging a combination of cross-branch spatial-temporal contrastive and projected feature contrastive loss functions, HSTRC can effectively extract robust representations from unlabeled time series data. Remarkably, by only fine-tuning the fully connected layers on top of extracted representations by HSTRC, we achieve the best performance across several Remaining Useful Life prediction datasets, showing up to 19.2% improvements over the state-of-the-art supervised learning methods and classical contrastive learning paradigms. Besides, further intensive experiments demonstrate HSTRC's effectiveness in active learning, out-of-distribution testing, and transfer learning scenarios.

Keywords: Time-Series Representation Learning · Contrastive Learning · RUL Prediction · Temporal-Spatial Contrasting

This work is supported within the LongLife (020E-100583532) project, funded by BMWK: German Federal Ministry for Economic Affairs and Climate Action.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78398-2_1.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15327, pp. 1–21, 2025.
https://doi.org/10.1007/978-3-031-78398-2_1

1 Introduction

The goal of Remaining Useful Life (RUL) prediction is extracting degradation representations from sensors multivariate time series (MTS) to estimate the time until a system can no longer maintain its normal operational state [21]. Accurate RUL predictions are expected to yield significant benefits, including the development of intelligent maintenance strategies, reduction in operational costs, and extension of equipment lifespan [29]. The degradation representations indicate the gradual decline of system performance and are mainly extracted by temporal and spatial features [8]. Recurrent Neural Networks (RNNs) and one-dimensional Convolutional Neural Networks (CNNs) are widely used to extract temporal features [12, 15, 24, 28], while Graph Neural Networks (GNNs) are utilized to learn spatial relationships among sensors [11, 21]. Recent studies [8, 29] demonstrate that RUL prediction accuracy can be improved by integrating these two types of features. Specifically, Huang et al. [8] introduced a cascaded architecture based on Graph Convolutional Networks (GCN) and Temporal Convolutional Networks (TCN), demonstrating significantly improved predictive accuracy compared to other existing methods. The above-mentioned research on RUL prediction employs supervised learning paradigms. As shown in Fig. 1a, the model learns hidden features \mathbf{H} from \mathbf{X} by corresponding RUL label y . However, in real-world scenarios, the scarcity of labeled MTS data reduces the efficiency of supervised learning paradigms.

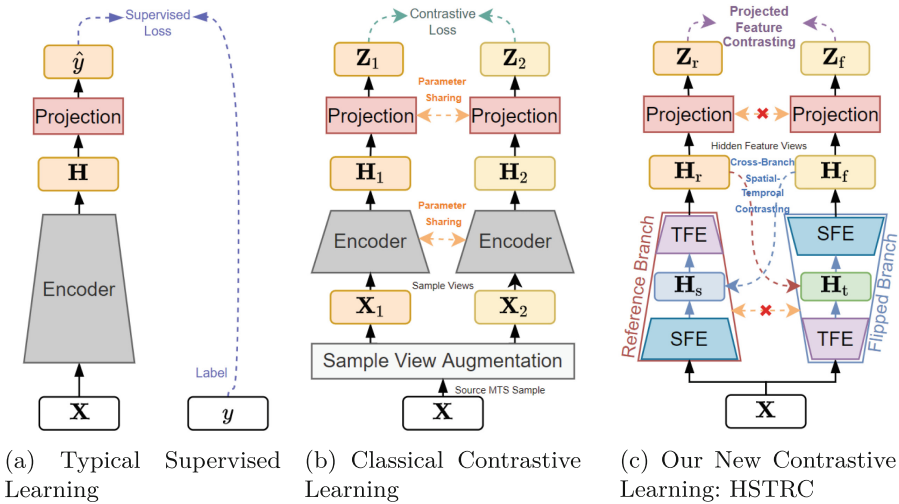


Fig. 1. Learning Paradigms Comparison

A potential solution is to employ the classical contrastive learning paradigm in Fig. 1b, as a subset of self-supervised learning, which extracts meaningful

representations from unlabeled MTS [4]. Most studies in recent years have utilized sample view augmentation techniques to create positive and negative pairs. As illustrated in Fig. 1b, a source MTS sample \mathbf{X} is augmented to two distinct sample views \mathbf{X}_1 and \mathbf{X}_2 , such as by adding noise [22], cropping [2], amplitude scaling [5], masking [5, 20], dropping [3, 26], and segmenting [27]. Then these two distinct sample views are fed into symmetric branches with shared parameters to extract representations \mathbf{Z}_1 and \mathbf{Z}_2 respectively. The classical loss function guides the model to minimize the distance between \mathbf{Z}_1 and \mathbf{Z}_2 [5] from the same source MTS sample \mathbf{X} . The sample views belonging to two different sources are commonly used as negative pairs [4, 5], while some studies [7, 27] use only positive pairs due to the existing periods in the MTS.

Above mentioned manual view augmentation techniques [2, 9, 20, 22, 26, 27] such as masking, adding noise, and shuffling randomly have applied in downstream tasks such as time series classification [3, 5, 22, 26] or forecasting [9, 23, 27]. However, when we face a new downstream task, such as RUL prediction, in which contrastive learning has not been widely applied, selecting an appropriate augmentation method without prior knowledge becomes subjective and challenging. Inappropriate random modifications to the original samples, such as adding excessive noise or altering the time series trend due to shuffling, can disrupt the latent patterns of the original MTS samples, resulting in suboptimal learned representations.

In response to these constraints, we propose an innovative contrastive learning paradigm referred to as Heterogeneous Spatio-Temporal Representation Contrasting (HSTRC). As shown in Fig. 1c, HSTRC differs from the classical contrastive learning paradigm by three main aspects:

- **Eliminating Sample View Augmentation:** Instead of sample view augmentation, HSTRC extracts distinct hidden feature views \mathbf{H}_r and \mathbf{H}_f from source MTS \mathbf{X} by two heterogeneous branches directly to avoid disturbing the latent pattern in MTS.
- **Heterogeneous Spatial-Temporal Flipped Branches:** Instead of shared symmetric branches, HSTRC constructs two heterogeneous branches with non-shared parameters by flipping the cascade sequence of the Spatial Feature Extractor (SFE) and the Temporal Feature Extractor (TFE). We define the SFE-TFE as the reference branch, extracting spatial features before temporal features, and TFE-SFE as the flipped branch, doing the reverse.
- **Joint Contrastive Loss Function:** HSTRC’s loss function combines cross-branch spatio-temporal contrasting and projected feature contrasting. The former encourages the two branches to cross-guide spatial and temporal feature learning. The latter aims to maximize the similarity between representations from the same source MTS while minimizing the similarity from different sources.

The main **Contributions** of this paper can be summarized as follows:

1. We propose a novel contrastive learning paradigm HSTRC, which is specific to MTS representation learning.

2. By only fine-tuning based on extracted representations, HSTRC achieves up to 19.2% improvement over the state-of-the-art supervised learning methods and classical contrastive learning paradigms in the RUL prediction scenario.
3. The further intensive experiments demonstrate that the learned representations by HSTRC are effective for downstream tasks under active learning, out-of-distribution (OOD) testing, and transfer learning settings.
4. We provide the code¹ to enable interested researchers to replicate our results and extend the application to other downstream tasks for MTS.

The overall structure and implementation details of HSTRC are described in Fig. 2 and Sec. 2, 3 and 4. We use RUL prediction as a downstream task in Sec. 5, and demonstrate the efficiency and accuracy of HSTRC on time series representation learning through intensive experiments in Sec. 6.

2 Overall Architecture

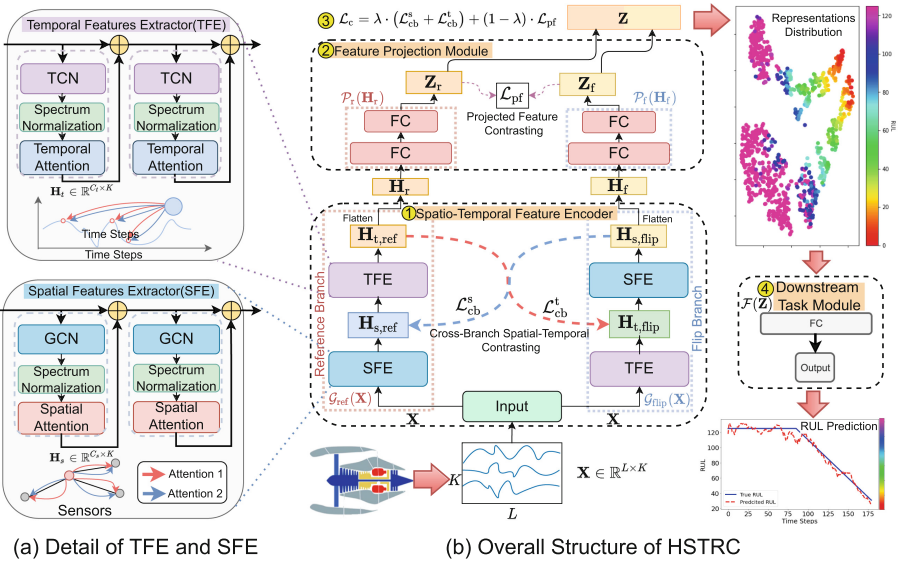


Fig. 2. HSTRC includes four primary steps: ① The spatial-temporal feature encoder with dual heterogeneous branches extracts distinct feature views \mathbf{H}_r and \mathbf{H}_f from source \mathbf{X} . ② The feature projection module further performs a nonlinear transformation on \mathbf{H}_r and \mathbf{H}_f . ③ Learning \mathbf{Z} by optimizing cross-branch spatial-temporal contrasting ($\mathcal{L}_{cb}^s, \mathcal{L}_{cb}^t$) and projected feature contrasting \mathcal{L}_{pf} loss functions. ④ Downstream task module for fine-tuning.

We define a sample of MTS as $\mathbf{X} = [\mathbf{x}_L^1, \mathbf{x}_L^2, \dots, \mathbf{x}_L^K] \in \mathbb{R}^{L \times K}$, where \mathbf{x}_L^k refers to univariate time series of the k -th sensor with length L , $k \in \{1, 2, \dots, K\}$. A

¹ <https://anonymous.4open.science/r/HSTRC/>

time series data set contains N instances is denoted as $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$. The overall architecture of the HSTRC framework, as illustrated in Fig. 2b, comprises four primary components:

- ① **Spatial-Temporal Feature Encoder $\mathcal{G}(\cdot)$** : $\mathcal{G}(\cdot)$ integrates two heterogeneous models without shared parameters, forming a dual-branch structure. One branch adopts the SFE-TFE cascading structure, denoted as the reference branch $\mathcal{G}_{\text{ref}}(\cdot)$, while the other represents a flip of the reference branch, i.e., TFE-SFE, defined as the flipped branch $\mathcal{G}_{\text{flip}}(\cdot)$. We generate two distinct hidden feature views $\mathbf{H}_r \in \mathbb{R}^{D_r}$ and $\mathbf{H}_f \in \mathbb{R}^{D_f}$ from the same source MTS \mathbf{X} as follows:

$$\{\mathbf{H}_r, \mathbf{H}_f\} = \mathcal{G}(\mathbf{X}) = \{\mathcal{G}_{\text{ref}}(\mathbf{X}), \mathcal{G}_{\text{flip}}(\mathbf{X})\}, \quad (1)$$

where D_r and D_f denotes the output dimension of the extracted views, with $D_r = D_f$.

- ② **Feature Projection Module $\mathcal{P}(\cdot)$** : $\mathcal{P}(\cdot)$ comprises two independent (i.e. no shared parameters) but homogeneous branches, $\mathcal{P}_r(\cdot)$ and $\mathcal{P}_f(\cdot)$, each consisting of two FC layers with spectrum normalization [17]. These two branches are utilized for the nonlinear transformation of \mathbf{H}_r and \mathbf{H}_f , yielding the projected features \mathbf{Z}_r and \mathbf{Z}_f , respectively. $\mathbf{Z}_r \in \mathbb{R}^D$ and $\mathbf{Z}_f \in \mathbb{R}^D$ are concatenated into $\mathbf{Z} \in \mathbb{R}^{2D}$, serving as representation of \mathbf{X} :

$$\mathbf{Z} = [\mathbf{Z}_r \parallel \mathbf{Z}_f] = [\mathcal{P}_r(\mathbf{H}_r) \parallel \mathcal{P}_f(\mathbf{H}_f)], \quad (2)$$

where D is the dimension of the projected features, and \parallel denotes the concatenation operation.

- ③ **Joint Contrastive Loss Function \mathcal{L}_c** : In the self-supervised learning phase, to learn robust representations \mathbf{Z} from the source MTS sample, we design a joint contrastive loss function \mathcal{L}_c , which comprises a cross-branch spatial $\mathcal{L}_{\text{cb}}^s$ and temporal $\mathcal{L}_{\text{cb}}^t$ contrasting in $\mathcal{G}(\cdot)$ and a projected feature contrasting \mathcal{L}_{pf} in $\mathcal{P}(\cdot)$. \mathcal{L}_c is presented as below:

$$\mathcal{L}_c = \lambda \cdot (\mathcal{L}_{\text{cb}}^s + \mathcal{L}_{\text{cb}}^t) + (1 - \lambda) \cdot \mathcal{L}_{\text{pf}}, \quad (3)$$

where λ is a hyperparameter representing the relative weight, with $0 \leq \lambda \leq 1$.

- ④ **Downstream Task Module $\mathcal{F}(\cdot)$** : $\mathcal{F}(\cdot)$ consists of a single FC layer. When training downstream tasks, the parameters of the spatio-temporal feature encoder $\mathcal{G}(\cdot)$ are frozen. Only the feature projection module $\mathcal{P}(\cdot)$ and the downstream task module $\mathcal{F}(\cdot)$ are fine-tuned through labeled samples. When employing RUL prediction as a downstream task, the goal is to establish a mapping function $y = \mathcal{F}(\mathbf{Z})$ that correlates a representation \mathbf{Z} with a corresponding RUL value $y \in \mathbb{R}$.

We will detail the implementation of the spatio-temporal feature encoder $\mathcal{G}(\cdot)$ in Sec. 3 and the contrastive loss function \mathcal{L}_c in Sec. 4.

3 Spatial-Temporal Feature Encoder

In this section, we describe the model implementation of SFE and TFE shown in Fig. 2a; and the Heterogeneous Spatial-Temporal Flipped Structure in Fig. 2b.

3.1 Spatial Feature Extractor

In SFE $\mathcal{S}(\cdot)$, we construct sensor graphs and extract spatial relationships between sensors from MTS through a GNN. As shown in Fig. 2a, the SFE consists of multiple residual-connected blocks, each of which consists of a GCN layer, a spectral normalization layer, and a spatial attention layer. The parameters of GCN are normalized through the spectrum normalization [17] layer to maintain distances between samples in the latent space. A M -head spatial attention layer [19] interprets sensor spatial relationships by enhancing attention between sensor connections.

3.2 Temporal Feature Extractor

Similarly, the TFE $\mathcal{T}(\cdot)$ shown in Fig. 2a includes multiple blocks with residual connections. Each block comprises a TCN layer, a spectral normalization layer, and a temporal attention layer. A M -head temporal attention mechanism is incorporated into each temporal block. Diverging from [8], which treats sensors equally, we improve the temporal attention mechanism in this paper to distinguish the significance of different sensors at each time step. For more implementation detail of SFE and TFE please see Appendix A. It is noted that SFE and TFE can also be replaced by other model structures.

3.3 Heterogeneous Spatial-Temporal Flipped Structure

As shown in Fig. 1c and Fig. 2b, this structure consists of two independent branches that do not share parameters. Due to their different cascading sequences, they are respectively referred to as the reference branch $\mathcal{G}_{\text{ref}}(\cdot)$ and the flipped branch $\mathcal{G}_{\text{flip}}(\cdot)$.

Reference Branch: The SFE in the reference branch is denoted as $\mathcal{S}_{\text{ref}}(\cdot)$, and the TFE as $\mathcal{T}_{\text{ref}}(\cdot)$. The sequence of cascading is SFE-TFE, meaning spatial features are first extracted from the MTS sample by the SFE $\mathbf{H}_{\text{s,ref}} = \mathcal{S}_{\text{ref}}(\mathbf{X}) \in \mathbb{R}^{C_s^{\text{ref}} \times K}$, which are then fed into the TFE to obtain the temporal features $\mathbf{H}_{\text{t,ref}} = \mathcal{T}_{\text{ref}}(\mathbf{H}_{\text{s,ref}}) \in \mathbb{R}^{C_t^{\text{ref}} \times K}$. The final temporal features are flattened to serve as the hidden feature view of the reference branch, defined as $H_r \in \mathbb{R}^{D_r}$, where $D_r = C_t^{\text{ref}} \cdot K$:

$$\mathbf{H}_r = \mathcal{G}_{\text{ref}}(\mathbf{X}) = \text{Flatten}(\mathcal{T}_{\text{ref}}(\mathcal{S}_{\text{ref}}(\mathbf{X}))). \quad (4)$$

Flipped Branch: Conversely, the flipped branch employs TFE $\mathcal{T}_{\text{flip}}(\cdot)$ and SFE $\mathcal{S}_{\text{flip}}(\cdot)$, with the flip cascading sequence, i.e. TFE-SFE. The flipped branch initially extracts temporal features through the TFE $\mathbf{H}_{\text{t,flip}} = \mathcal{T}_{\text{flip}}(\mathbf{X}) \in \mathbb{R}^{C_t^{\text{flip}} \times K}$, which are then input to the SFE $\mathbf{H}_{\text{s,flip}} = \mathcal{S}_{\text{flip}}(\mathbf{H}_{\text{t,flip}}) \in \mathbb{R}^{C_s^{\text{flip}} \times K}$ to extract spatial features. It is noteworthy that $C_s^{\text{ref}} = C_t^{\text{flip}}$ and $C_t^{\text{ref}} = C_s^{\text{flip}}$. Ultimately, the spatial features are flattened as the output view of the flipped branch, defined as $\mathbf{H}_f \in \mathbb{R}^{D_f}$, where $D_f = C_s^{\text{flip}} \cdot K$:

$$\mathbf{H}_f = \mathcal{G}_{\text{flip}}(\mathbf{X}) = \text{Flatten}(\mathcal{S}_{\text{flip}}(\mathcal{T}_{\text{flip}}(\mathbf{X}))). \quad (5)$$

\mathbf{H}_r and \mathbf{H}_f are two distinct spatial-temporal feature views of heterogeneous dual branches from the same source MTS \mathbf{X} . These hidden feature views are then directed into the feature projection module to obtain the representation \mathbf{Z} . The visualizations of two branches' attention from same \mathbf{X} are located in Appendix C.

4 Joint Contrastive Loss Function

HSTRC is jointly optimized through two types of contrastive loss functions in the self-learning phase. These two losses are tailored for the spatial-temporal feature encoder $\mathcal{G}(\cdot)$ and the feature projection moduler $\mathcal{P}(\cdot)$, respectively.

4.1 Cross-Branch Spatial-Temporal Contrasting

This loss function is implemented within the spatial-temporal feature encoder. Given the b -th MTS sample from a batch with size B , we extract spatial hidden features ($\mathbf{H}_{s,\text{ref}}^b$ and $\mathbf{H}_{s,\text{flip}}^b$) as well as temporal features ($\mathbf{H}_{t,\text{ref}}^b$ and $\mathbf{H}_{t,\text{flip}}^b$) through the reference branch and flipped branch. Taking the spatial contrasting as an example, we first flatten two spatial features as $\overline{\mathbf{H}}_{s,\text{ref}}^b$ and $\overline{\mathbf{H}}_{s,\text{flip}}^b$ and then apply a bilinear function to maintain the mutual information between the inputs, which is $\exp(\mathcal{K}_s(\overline{\mathbf{H}}_{s,\text{flip}}^b)(\overline{\mathbf{H}}_{s,\text{ref}}^b)^T)$. Here, $\mathcal{K}_s(\cdot)$ is a linear mapping function that projects $\overline{\mathbf{H}}_{s,\text{flip}}^b$ to the same dimensional space as $\overline{\mathbf{H}}_{s,\text{ref}}^b$, defined as $\mathcal{K}_s(\cdot) : \mathbb{R}^{C_s^{\text{flip}} \cdot K} \rightarrow \mathbb{R}^{C_s^{\text{ref}} \cdot K}$. Specifically, we transform the $\overline{\mathbf{H}}_{s,\text{flip}}^b$ into same shape with the $\overline{\mathbf{H}}_{s,\text{ref}}^b$. The contrastive loss $\mathcal{L}_{\text{cb}}^s$ defined in Eq. (6) aims to maximize the dot product between $\overline{\mathbf{H}}_{s,\text{flip}}^b$ and $\overline{\mathbf{H}}_{s,\text{ref}}^b$ from the same source MTS sample, while minimizing the dot product with spatial feature views generated from other samples within the same batch. This encourages the contrastive target to align the transformed $\mathcal{K}_s(\overline{\mathbf{H}}_{s,\text{flip}}^b)$ with its corresponding feature $\overline{\mathbf{H}}_{s,\text{ref}}^b$. A similar loss function $\mathcal{L}_{\text{cb}}^t$ in Eq. (7) is also formulated for the temporal contrasting:

$$\mathcal{L}_{\text{cb}}^s = -\frac{1}{B} \sum_{b=1}^B \log \frac{\exp(\mathcal{K}_s(\overline{\mathbf{H}}_{s,\text{flip}}^b)(\overline{\mathbf{H}}_{s,\text{ref}}^b)^T)}{\sum_{j \in B} \exp(\mathcal{K}_s(\overline{\mathbf{H}}_{s,\text{flip}}^b)(\overline{\mathbf{H}}_{s,\text{ref}}^j)^T)}, \quad (6)$$

$$\mathcal{L}_{\text{cb}}^t = -\frac{1}{B} \sum_{b=1}^B \log \frac{\exp(\mathcal{K}_t(\overline{\mathbf{H}}_{t,\text{ref}}^b)(\overline{\mathbf{H}}_{t,\text{flip}}^b)^T)}{\sum_{j \in B} \exp(\mathcal{K}_t(\overline{\mathbf{H}}_{t,\text{ref}}^b)(\overline{\mathbf{H}}_{t,\text{flip}}^j)^T)}. \quad (7)$$

4.2 Projected Feature Contrasting

In the feature projection module, we utilize contrastive learning for projected features to explore discriminative representations. Within a batch containing B original MTS samples, the b -th sample generates two projected features, namely \mathbf{Z}_r^b and \mathbf{Z}_f^b , resulting in a total of $2B$ projected feature views. For each \mathbf{Z}_r^b , its corresponding \mathbf{Z}_f^b is considered a positive sample from the same source MTS, thereby making a positive pair $(\mathbf{Z}_r^b, \mathbf{Z}_f^b)$. Concurrently, the remaining $(2B - 2)$ feature views from other samples in the same batch serve as negative samples of \mathbf{Z}_r^b , leading to $(2B - 2)$ negative pairs for \mathbf{Z}_r^b . Based on this setup, we define the projected feature contrastive loss \mathcal{L}_{pf} in Eq. (8). Specifically, given \mathbf{Z}_r^b , we divide its similarity with the positive sample \mathbf{Z}_f^b by the sum of its similarities with all other $(2B - 1)$ projected feature views in the batch, including a positive pair and $(2B - 2)$ negative pairs. The goal of \mathcal{L}_{pf} is maximizing the similarity between positive pairs while minimizing the similarity between negative pairs:

$$\mathcal{L}_{\text{pf}} = -\frac{1}{B} \sum_{b=1}^B \log \frac{\exp(\text{sim}(\mathbf{Z}_r^b, \mathbf{Z}_f^b)/\tau)}{\sum_{i=1}^B 1_{[i \neq b]} \exp(\text{sim}(\mathbf{Z}_r^b, \mathbf{Z}_r^i)/\tau) + \sum_{i=1}^B \exp(\text{sim}(\mathbf{Z}_r^b, \mathbf{Z}_f^i)/\tau)}, \quad (8)$$

where $\text{sim}(\mathbf{a}, \mathbf{b})$ is denoted as the cosine similarity between vectors \mathbf{a} and \mathbf{b} , $1_{[i \neq b]} \in 0, 1$ is a mask function that equals 1 if $i \neq b$, and τ represents the temperature parameter. The joint contrastive loss function \mathcal{L}_c represents in Eq. (3) is a combination of cross-branch spatial-temporal contrasting \mathcal{L}_{cb} and projected feature contrasting \mathcal{L}_{pf} .

5 Experimental Setup

5.1 Datasets Description

We utilize RUL prediction as the downstream task to verify the performance of HSTRC. For a fair comparison with existing work, we conduct our experiments based on the widely recognized Commercial Modular Aerospace Propulsion System Simulation (C-MAPSS) data collection [18]. As detailed in Table. 1, this data collection comprises four datasets referred to as from FD001 to FD004 respectively, and used as a benchmark for numerous research articles concerned with RUL prediction [1, 8, 10, 12–15, 21, 24, 29]. Each dataset contains multiple engine records and is divided into a training and a test set. Each engine recorded time series with 24 sensors (symbolizing $K = 24$ in \mathbf{X}) and corresponding operational duration denoted as RUL. The measurements of engines are collected under different operational conditions. FD001 and FD003 were measured by single conditions, while FD002 and FD004 contain the measurements collected under six operational conditions, which present more complex challenges for the accurate RUL prediction [8]. Consistent with other RUL prediction research [6, 8, 25], we normalize the data based on operating conditions and use sliding windows with length 50 (i.e. $L = 50$ in \mathbf{X}) to segment the raw time series data. The operating conditions and sensor measurements constituted the final input vectors for the models.

5.2 Evaluation Metrics and Implementation Details

Following previous works [1, 12, 15, 21, 24], we utilize the RMSE and Score [12] to evaluate the performance of RUL prediction. The lower RMSE and Score indicate higher prediction accuracy. For more details of metrics, see Appendix B. The spatial-temporal feature encoder extracts features with the dimension $D_r = D_f = 240$, and the output dimension of each branch in the feature projection module is $D = 50$. The number of attention heads M is 4 and temperature τ is 0.2. We employ the Adam optimizer with a learning rate of 0.0001 for FD002 and FD004, 0.005 for FD001 and FD003, a batch size of 50, and training epochs of 32. Each experiment is repeated five times with different random seeds to ensure reliability. The mean values of relevant evaluation metrics are computed for performance assessment.

6 Experimental Results

In this section, we conduct five experiments. Firstly, the effectiveness of HSTRC is validated in Sec. 6.1. Subsequently, Sec. 6.2 utilizes RUL prediction as a downstream task to compare HSTRC with other existing methods. Sec. 6.3 and 6.4 explore HSTRC’s representation capabilities under active learning, OOD testing, and transfer learning settings. Lastly, we employ model ablation study in Sec. 6.5 to analyze the impact of spatial-temporal feature encoder on prediction results.

Table 1. Statistics of the C-MAPSS dataset

Subsets	FD001	FD002	FD003	FD004
Number of Training Engines	100	260	100	249
Number of Test Engines	100	259	100	248
Number of Operation Conditions	1	6	1	6

6.1 Model Effectiveness and Sensitivity Analysis

This experiment aims to: (1) Validate the effectiveness of HSTRC; (2) Analyze the impact of the hyperparameter λ on the loss function Eq. (3). HSTRC learns the hidden degradation representation \mathbf{Z} from unlabeled MTS sample \mathbf{X} . Note that HSTRC can not access the RUL labels during the self-supervised learning phase.

Fig. 3a shows the distribution of \mathbf{Z} without using the joint contrastive loss function \mathcal{L}_c , where the representations are randomly distributed, failing to distinguish degradation information. After introducing \mathcal{L}_c , corresponding to Fig. 3b and 3c, the representations demonstrate clustering and stratification effects in

accordance with RUL decreases. Fig. 3b shows the representations distribution when $\lambda = 0$, i.e., only using the projected feature contrasting \mathcal{L}_{pf} . When $\lambda = 1$, we solely use the cross-branch spatio-temporal contrasting \mathcal{L}_{cb} , with the distribution shown in Fig. 3c. The comparison between Fig. 3b and 3c indicates that \mathcal{L}_{pf} primarily clusters similar samples together, while \mathcal{L}_{cb} further increases the distance between samples with different degradation levels. The above observations confirm that our proposed loss function \mathcal{L}_c can help HSTRC effectively extract degradation information from unlabeled data through contrastive learning.

Fig. 3d quantifies the impact of λ on downstream task performance in terms of RMSE and Score. The HSTRC exhibits robustness to variations in λ from 0.2 to 0.8. More visualization results such as attention in the dual branches can be found in Appendix C.

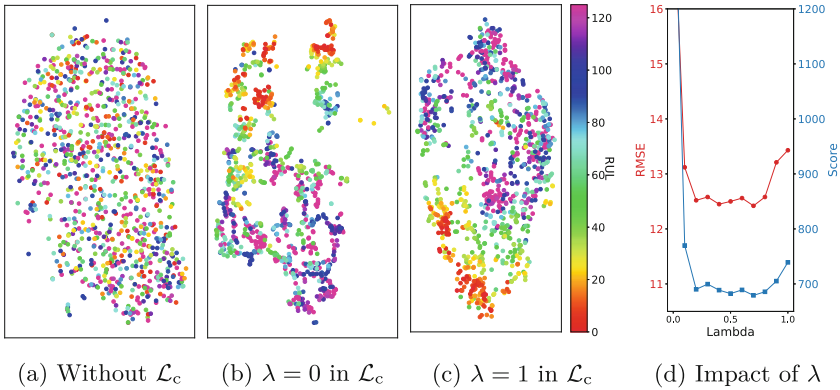


Fig. 3. Sensitivity Analysis of λ . (a), (b) and (c) visualize the distribution of HSTRC’s latent representations \mathbf{Z} regarding the first 1000 samples in the FD002 dataset using t-SNE. Each circle represents a MTS sample, with its color indicating the sample’s actual RUL value. (d) presents the RMSE and Score on the FD002 under the different λ values.

6.2 Comparison with State-of-the-Art

This experiment compares HSTRC with existing research in RUL prediction. As shown in Table 2, all existing approaches utilize end-to-end supervised learning methods. We cite the results from their original papers. To explore the impact of sample view augmentation on distorting degradation information in MTS, we also adopt the advanced method TS-TCC [5] based on the classical contrastive learning paradigm to RUL prediction. TS-TCC augments source samples by adding noise and shuffling. Additionally, we create a baseline that does not employ the contrastive loss function L_c to compare the performance of HSTRC.

By solely fine-tuning the FC layers, HSTRC delivers not only the best performance on FD001, FD002, and FD004 but also near state-of-the-art performance

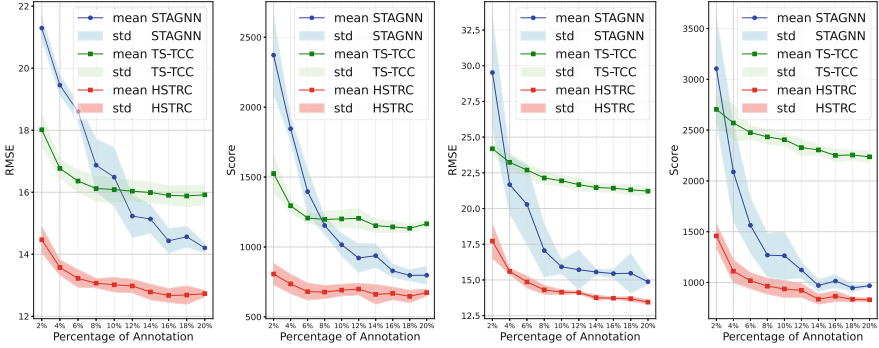
Table 2. Performance Comparison in RUL Prediction

Learning Type	Datasets	FD001		FD002		FD003		FD004	
		Methods	RMSE	Score	RMSE	Score	RMSE	Score	RMSE
Supervised	HALSTM [1]	14.53	322.4	N/A	N/A	N/A	N/A	27.08	5649.1
	RNN [24]	13.58	228	19.59	2650	19.16	1727	22.15	2901
	AGCNN [15]	12.42	225.5	19.43	1492	13.39	227	21.50	3392
	GCN [21]	12.76	266	N/A	N/A	12.07	278	N/A	N/A
	LSTMBMS [13]	14.89	481.1	26.86	7982	15.11	493.4	27.11	5200
	DLSTM [16]	12.29	N/A	17.87	N/A	14.34	N/A	21.81	N/A
	HAGCN [10]	11.93	222.3	15.05	1144.1	11.53	240.3	15.74	1218.6
	BDL [14]	18.60	2774	22.90	7734	27.90	19990	28.10	53295
	GAT [11]	13.21	303.1	17.25	5338.8	15.36	507.5	21.44	2971.9
	RGCNU [29]	11.18	173.5	16.22	1148.16	11.52	225.0	19.11	2215.9
STAGNN [8]	11.50	187.2	13.81	826.3	11.05	196.0	14.30	1038.5	
Self-Super. +	TS-TCC [5]	14.51	358.7	16.25	1356.3	18.39	1142.9	20.65	2350.6
	Baseline	28.36	1959.3	43.45	119202	31.76	8371.3	38.81	61777
Fine-Tuning	HSTRC	11.10	167.7	12.42	679.2	11.51	226.0	13.33	838.1

on FD003. Furthermore, compared with the cutting-edge supervised learning model STAGNN [8], HSTRC archives up to 19.2% improvement in Score. These results demonstrate that self-supervised learning paradigms can be successfully applied to RUL prediction, and validate that our proposed HSTRC can effectively learn representations from unlabeled MTS datasets. The visualized representation distribution before and after fine-tuning can be found in Appendix D. Comparative results between HSTRC and TS-TCC validate the hypothesis in Sec. 1 that sample view augmentations could disturb the pattern of MTS, leading to suboptimal performance. Compared to the baseline in Table. 2, HSTRC demonstrates significant improvements in RMSE and Score on all datasets. The results of the baseline are consistent with our observations in Fig. 3a that the baseline is not able to learn effective degradation information without L_c .

6.3 Active Learning Experiment

Active learning aims to approximate the efficacy of fully supervised learning by selectively annotating a minimal subset of samples from a large pool of unlabeled data. This study examines the performance of HSTRC and TS-TCC in integrating self-supervised and active learning. Initially, HSTRC and TS-TCC extract high-dimensional representations from unlabeled samples. Subsequently, a limited set of these samples is annotated for fine-tuning of FC layers on the top. To provide a baseline, the state-of-the-art supervised model STAGNN is utilized for comparative analysis, undergoing end-to-end training in the active

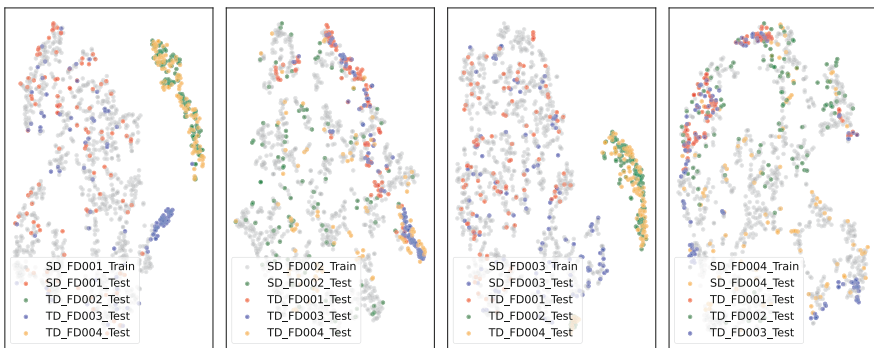


(a) RMSE on FD002 (b) Score on FD002 (c) RMSE on FD004 (d) Score on FD004

Fig. 4. Active Learning with Random Sampling on FD002 and FD004

learning scenario. The process involves executing 10 active learning cycles, where 2% of the unlabeled samples are annotated per cycle through random sampling.

Under the limited annotated samples, the performance of STAGNN and TS-TCC (indicated by the blue and green lines in Fig.4) is suboptimal, whereas our HSTRC, through only fine-tuning FC layers, achieved better performance on both the RMSE and Score (the red line in Fig.4), with a maximum improvement of 66.1%. For the FD002 and FD004 datasets, by annotating solely 16% and 20% of the samples respectively, HSTRC can match the performance of fully fine-tuning with 100% labeled data in Table. 2. This highlights that HSTRC leveraging robust representation by contrastive learning can enhance the efficiency in active learning scenarios.



(a) FD001 as SD (b) FD002 as SD (c) FD003 as SD (d) FD004 as SD

Fig. 5. OOD Testing. The gray circles show the distribution of SD training sets. Circles in other colors are representations of TD samples extracted by HSTRC after self-supervised training on SD. Their colors correspond to their datasets.

6.4 OOD Testing and Transfer Learning Experiment

The goal of transfer learning is to apply the knowledge from the source domain (SD) to the target domain (TD), under the premise that the data distributions between the two domains are similar. OOD testing can assess whether the SD and TD are aligned in an identical distribution. This experiment aims to evaluate whether HSTRC can accurately identify OOD samples and perform transfer learning effectively. Specifically, we alternately choose the FD001-FD004 datasets as the SD and the remaining three test sets as the TD. The normalization scaler from the SD is applied to the TD data. After HSTRC is trained based on the SD, we transfer the learned knowledge to the TD.

Fig. 5 shows the OOD testing results with FD001-FD004 as SD, respectively. Overall, the distributions of FD001 and FD003 are close to each other, indicating they might have been collected under similar operating conditions. A similar distribution also exists between FD002 and FD004. When FD001 and FD003 serve as SD (as shown in Fig. 5a and 5c), FD002 and FD004 are considered as OOD data because FD001 and FD003 were solely collected under a single condition, whereas FD002 and FD004 encompass six different operating conditions, making it challenging for FD001 and FD003 to adapt to the diversity. Conversely, when FD002 and FD004 serve as SD (as shown in Fig. 5b and 5d), FD001 and FD003 appear as in-distribution data, with their representations primarily located in the edge of the SD distribution. This suggests that FD002 and FD004 could transfer their knowledge to FD001 and FD003.

Table 3. Performance of HSTRC for Transfer Learning

		Target		FD001		FD002		FD003		FD004	
		Source	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score	
Zero-Shot	FD001	-	-	51.07	35645	21.69	1861.0	54.23	45332		
	FD002	16.70	336.7	-	-	19.07	684.8	18.27	2292.7		
	FD003	12.19	185.2	46.23	23281	-	-	47.71	23394		
	FD004	14.07	285.7	13.42	809.8	13.66	288.4	-	-		
Fine Tuning	FD001	-	-	41.68	32299	11.54	232.1	43.48	37618		
	FD002	11.18	182.8	-	-	10.61	182.3	13.53	956.96		
	FD003	11.58	199.0	42.64	36521	-	-	43.45	26901		
	FD004	11.33	188.1	13.39	808.8	11.28	194.7	-	-		

Table 3 presents the quantitative results of applying HSTRC to transfer learning. We employ two transfer learning strategies: Zero-Shot and Fine-Tuning. The former directly applies the model trained on the SD to the TD, while the latter further fine-tunes the downstream task through labeled TD data. The quantitative outcomes are consistent with the observations from the OOD tests in Fig. 5, indicating effective transfer learning between FD001 and FD003, as well

as between FD002 and FD004. The transfer performance from FD001 and FD003 to FD002 and FD004 showed poor results due to the latter being OOD for the former. In contrast, the knowledge learned from FD002 and FD004 could be effectively transferred to FD001 and FD003. Even with Zero-Shot, the performance is comparable to or even better than most existing fully supervised methods in Table 2. This experiment not only proves HSTRC’s ability to identify OOD samples accurately but also demonstrates its effectiveness in performing transfer learning between SD and TD with similar data distributions.

Table 4. Ablation Analysis of Spatial-Temporal Feature Encoder

Approaches	FD001		FD002		FD003		FD004	
	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score
STEF	11.82	211.3	14.55	868.9	12.64	284.1	15.97	1202.5
STFE_RS	11.71	187.3	12.76	814.7	11.72	237.5	14.55	955.8
STFE_A	13.58	228	20.65	2892.6	12.26	259.0	14.78	1128.0
STFE_AR	11.50	189.3	13.14	704.3	12.71	264.8	17.20	1470.3
STFE_AS	20.51	2739.8	20.07	2307.3	22.42	1520.6	22.64	3375.7
STFE_ARS	11.10	167.7	12.42	679.2	11.51	226.0	13.33	838.1

6.5 Model Ablation Analysis

Compared to the existing model STAGNN [8], we incorporate improvements in the spatial-temporal feature encoder of HSTRC, including the enhancement of the temporal attention mechanism, spectral normalization, and residual connections. This experiment aims to conduct an ablation analysis of these three enhancements to assess their impact on the performance of downstream tasks. Therefore, we decomposed the spatial-temporal feature encoder (STFE_AS) into five submodels: STFE contains only GCN-TCN cascade structure; STFE_A introduces the attention mechanism; STFE_SR adds spectral normalization and residual connections; and STFE_AR and STFE_AS are based on STFE_A but incorporate residual connections and spectral normalization, respectively.

Table 4 shows the results of the ablation analysis across four datasets. Compared to the STFE, STFE_SR and STFE_A showed up to a 16.4% improvement in Score. This indicates that spectral normalization, residual connections, and attention effectively prevent feature collapse, enabling the feature encoder to extract more diverse features. Consider STFE_AS and STFE_ARS, only combining attention and spectral normalization will make the results less optimized, but further adding residual connections can achieve the best performance. Appendix E shows an additional explanation and visualization of submodels’ representations.

7 Conclusion and Future Work

We propose a novel contrastive learning paradigm for MTS termed HSTRC and introduce contrastive learning to RUL prediction for the first time. Unlike classical paradigms that manipulate the original MTS in sample view augmentation, HSTRC uses dual branches with a heterogeneous spatial-temporal flipped structure to generate two distinct feature views from the same source without any disturbance. Integrating cross-branch spatial-temporal contrastive with projected feature contrastive loss functions, HSTRC efficiently extracts robust features from unlabeled MTS. Only fine-tuning the FC layers on the top, HSTRC achieves superior performance on several RUL prediction datasets, with up to a 19.2% improvement over the state-of-the-art supervised learning methods and classical contrastive learning paradigms. Additionally, HSTRC demonstrates its effectiveness in active learning scenarios, achieving close fully supervised performance with only 20% of the labeled samples. It also accurately identifies OOD data between SD and TD, providing valuable insights for transfer learning. This paper mainly focuses on RUL prediction as a use case, but HSTRC as a novel contrastive learning paradigm holds potential for other downstream tasks, such as time series classification and anomaly detection, which we aim to explore in future work.

A Implementation Detail of Spatial and Temporal Feature Extractor

A.1 Spatial Feature Extractor

In SFE, we construct sensor graphs and extract spatial relationships between sensors from MTS through GNN. A graph structure symbolizes sensors as nodes encapsulated in an adjacency matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$. The learnable adjacency matrix represent as a parameter matrix $\Theta \in \mathbb{R}^{K \times K}$, activated by the Tanh function: $\mathbf{A} = \text{Tanh}(\Theta)$. As shown in Fig. 2a, the SFE consists of multiple residual-connected blocks, each of which consists of a GCN layer, a spectral normalization layer, and a spatial attention layer. The parameters of GCN are normalized through the spectrum normalization [17] layer to maintain distances between samples in the latent space.

A multi-head spatial attention layer [19] interprets sensor spatial relationships by enhancing attention between sensor connections. The input of spatial attention layer, $\hat{\mathbf{H}}_s \in \mathbb{R}^{C_s \times K}$, represents the spectral normalized feature extracted by GCN, where $C_s \times K$ is the output shape after GCN convolution and $\hat{\mathbf{H}}_s^i \in \mathbb{R}^{C_s}$ corresponding to the i -th sensor. The output of the multi-head spatial attention layer for the i -th sensor, \mathbf{H}_s^i , is given by:

$$\mathbf{H}_s^i = \frac{1}{M} \sum_{m=1}^M \sum_{j \in \mathbb{N}_i} \alpha_{ij}^m \hat{\mathbf{H}}_s^j, \quad (9)$$

with M being the number of attention heads, α_{ij}^m being the attention coefficient for the m -th head and \mathbb{N}_i denotes the neighbors of sensor i .

A.2 Temporal Feature Extractor

Similarly, the TFE includes multiple blocks with residual connections. Each block comprises a TCN layer, a spectral normalization layer, and a temporal attention layer. The spectral normalized output of TCN is denoted as $\hat{\mathbf{H}}_t \in \mathbb{R}^{C_t \times K}$. A multi-head temporal attention mechanism is incorporated into each temporal block. Diverging from [8], which treated sensors equally, we propose a novel mechanism in this paper to enable distinguishing the significance of different sensors at each time step. The matrix $\beta \in \mathbb{R}^{C_t \times K}$, representing the significance of each sensor across C_t steps, is calculated as follows:

$$\beta = \frac{\exp\left(\text{Sigmod}\left(\hat{\mathbf{H}}_t \mathbf{W}_t + \mathbf{b}\right)\right)}{\sum \exp\left(\text{Sigmod}\left(\hat{\mathbf{H}}_t \mathbf{W}_t + \mathbf{b}\right)\right)}, \quad (10)$$

where $\mathbf{W}_t \in \mathbb{R}^{K \times K}$ and \mathbf{b} are the weight and bias of the attention layer, respectively. The output of the multi-head temporal attention layer is:

$$\mathbf{H}_t = \frac{1}{M} \sum_{m=1}^M \hat{\mathbf{H}}_t \odot \beta^m. \quad (11)$$

Here, M represents the number of attention heads, β^m is the attention coefficient matrix for the m -th head, and the \odot symbol denotes element-wise multiplication.

B Evaluation Metrics

The RMSE and Score are utilized in the experiments. RMSE is a classic metric that measures the error between the actual and predicted values in the regression task:

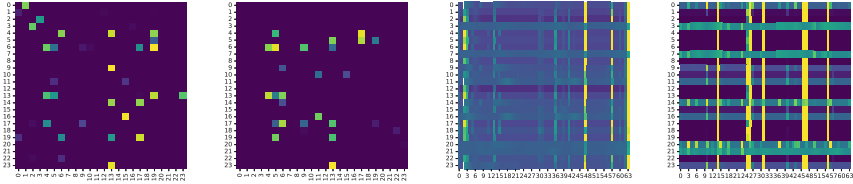
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}, \quad (12)$$

where N is the total number of predicted samples, y_i represents the actual RUL value of the i -th sample and \hat{y}_i is the predicted RUL value. Unlike RMSE, which can not distinguish between early and delayed predictions, the Score will bring more penalties to delayed RUL predictions. The lower Score indicates high prediction accuracy.

$$\text{Score} = \begin{cases} \sum_{i=1}^N (e^{-\frac{\hat{y}_i - y_i}{13}} - 1), & \text{if } \hat{y}_i < y_i \\ \sum_{i=1}^N (e^{\frac{\hat{y}_i - y_i}{10}} - 1), & \text{if } \hat{y}_i \geq y_i \end{cases} \quad (13)$$

C Attention Visualization of Heterogeneous Branches

In Sec.6.1, we validated the effectiveness of HSTRC in extracting features through heterogeneous dual branches combined with our proposed contrastive objective function. As a novel contrastive learning paradigm, HSTRC is based on the following two core hypotheses:



(a) Spatial Attention Reference Branch (b) Spatial Attention Flip Branch (c) Temporal Attention Reference Branch (d) Temporal Attention Flip Branch

Fig. 6. Spatial-Temporal Attention on Reference and Flip Branches

1. Heterogeneous models can learn different hidden feature views from the same source data. Unlike existing contrastive learning methods that use homogeneous models with shared parameters to contrast different augmented sample views, our paradigm employs two independent heterogeneous models.
2. Changing the order of spatio-temporal feature extraction can produce different hidden feature views. Current state-of-the-art RUL prediction methods [8] typically extract spatial features from the original time series first and then extract temporal features (SFE-TFE). By reversing this order of spatio-temporal feature extraction, i.e. extracting temporal features before spatial features (TFE-SFE), we create heterogeneous models that can generate different feature views.

To validate these assumptions, an intuitive method is to visualize the spatio-temporal attentions of the two branches based on the same source MTS. Spatial attention reflects the importance of sensor connections, while temporal attention can reveal the importance of different sensors at each time step. The comparison of attention of the reference and the flipped branch on the same source data, as shown in Fig. 6, indicates that heterogeneous branches focus on different aspects of the same source data. This difference leads to the generation of different hidden feature views, thus validating our hypotheses and further proving the effectiveness of our proposed HSTRC.

D Distribution of Representation after Fine-Tuning

When training downstream tasks in Sec. 6.2, the parameters of the spatio-temporal feature encoder $\mathcal{G}(\cdot)$ are frozen. Only the feature projection module $\mathcal{P}(\cdot)$ and the downstream task module $\mathcal{F}(\cdot)$ are fine-tuned through supervised learning, i.e., using labeled data. In this section, we compare the distribution changes of representations before and after fine-tuning. According to Fig. 7, after fine-tuning through supervised learning, the distribution of representations becomes more concentrated compared to before fine-tuning, while the overall trend remains. This validates the robustness of the representations learned during self-supervised learning phase.

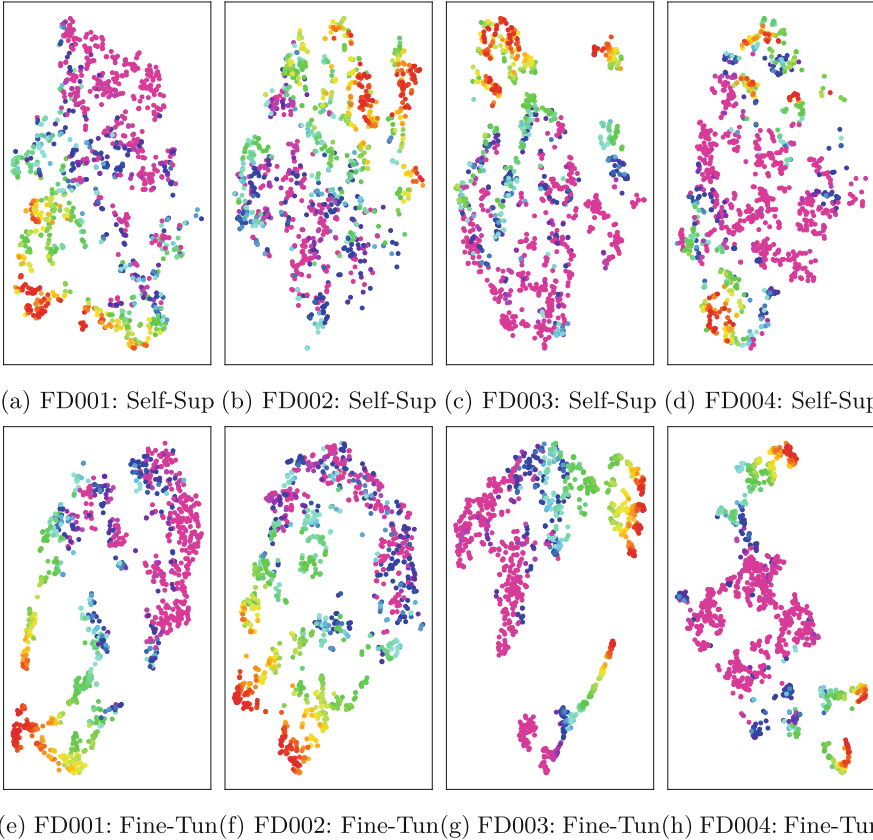


Fig. 7. The representation distribution before and after fine-tuning for downstream task.

E Distribution of Representation for Ablation Analysis

In Sec. 6.5, we decomposed the spatio-temporal feature encoder (STFE_ASR) into five submodules for ablation analysis: STFE, STFE_A, STFE_SR, STFE_AR, and STFE_AS. We conducted a quantitative result analysis in Table. 4. This section visualizes the representation distribution of different mechanisms during the self-supervised learning phase.

Based on the visualization results for STFE_A and STFE_AS in Fig. 8, we found that the representations using attention mechanisms not only distribute according to the RUL trend but also cluster together based on the same engine. This visualization potentially suggests that STFE_A and STFE_AS may achieve the best performance on downstream tasks. However, contrary to intuition, the result in Table. 4 shows that STFE_A and STFE_AS perform the worst in FD001 to FD004. This indicates that without residual connections, the robustness of

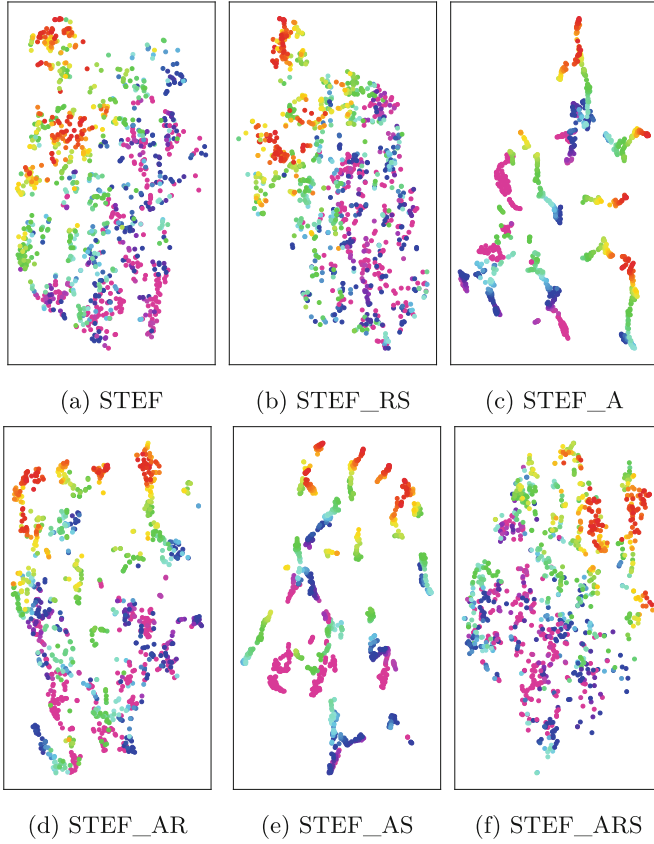


Fig. 8. The representation distribution before and after fine-tuning

the features learned in self-supervised learning is poor, which is detrimental to fine-tuning for downstream tasks.

References

1. Chen, Z., Wu, M., Zhao, R., Guretno, F., Yan, R., Li, X.: Machine remaining useful life prediction via an attention-based deep learning approach. *IEEE Trans. Industr. Electron.* **68**(3), 2521–2531 (2020)
2. Choi, H., Kang, P.: Multi-task self-supervised time-series representation learning. *arXiv preprint [arXiv:2303.01034](https://arxiv.org/abs/2303.01034)* (2023)
3. Dai, Y., Mei, Z., Li, J., Li, Z., Wei, K., Ding, M., Guo, S., Chen, W.: Clustering-based contrastive learning for fault diagnosis with few labelled samples. *IEEE Transactions on Instrumentation and Measurement* (2023)
4. Deldari, S., Xue, H., Saeed, A., He, J., Smith, D.V., Salim, F.D.: Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv preprint [arXiv:2206.02353](https://arxiv.org/abs/2206.02353)* (2022)

5. Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwok, C.K., Li, X., Guan, C.: Time-series representation learning via temporal and contextual contrasting. In: International Joint Conference on Artificial Intelligence, IJCAI (2021)
6. Heimes, F.O.: Recurrent neural networks for remaining useful life estimation. In: 2008 international conference on prognostics and health management. pp. 1–6. IEEE (2008)
7. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6535–6545 (2021)
8. Huang, Z., He, Y., Sick, B.: Spatio-temporal attention graph neural network for remaining useful life prediction. In: Computational Science and Computational Intelligence. IEEE (2023)
9. Li, L., Yang, K., Luo, F., Bi, J.: Sts-ccl: Spatial-temporal synchronous contextual contrastive learning for urban traffic forecasting. International Conference on Acoustics, Speech, & Signal Processing (ICASSP) (2024)
10. Li, T., Zhao, Z., Sun, C., Yan, R., Chen, X.: Hierarchical attention graph convolutional network to fuse multi-sensor signals for remaining useful life prediction. Reliability Engineering & System Safety **215**, 107878 (2021)
11. Li, T., Zhou, Z., Li, S., Sun, C., Yan, R., Chen, X.: The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study. Mech. Syst. Signal Process. **168**, 108653 (2022)
12. Li, X., Ding, Q., Sun, J.Q.: Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety **172**, 1–11 (2018)
13. Liao, Y., Zhang, L., Liu, C.: Uncertainty prediction of remaining useful life using long short-term memory network based on bootstrap method. In: 2018 IEEE international conference on prognostics and health management (icphm). pp. 1–8. IEEE (2018)
14. Lin, Y.H., Li, G.H.: A bayesian deep learning framework for rul prediction incorporating uncertainty quantification and calibration. IEEE Trans. Industr. Inf. **18**(10), 7274–7284 (2022)
15. Liu, H., Liu, Z., Jia, W., Lin, X.: Remaining useful life prediction using a novel feature-attention-based end-to-end approach. IEEE Trans. Industr. Inf. **17**(2), 1197–1207 (2020)
16. Miao, H., Li, B., Sun, C., Liu, J.: Joint learning of degradation assessment and rul prediction for aeroengines via dual-task deep lstm networks. IEEE Trans. Industr. Inf. **15**(9), 5023–5032 (2019)
17. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. International Conference on Learning Representations (ICLR) (2018)
18. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 international conference on prognostics and health management. pp. 1–9. IEEE (2008)
19. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al.: Graph attention networks. stat **1050**(20), 10–48550 (2017)
20. Wang, L., Bai, L., Li, Z., Zhao, R., Tsung, F.: Correlated time series self-supervised representation learning via spatiotemporal bootstrapping. International Conference on Automation Science and Engineering (CASE) (2023)
21. Wang, M., Li, Y., Zhang, Y., Jia, L.: Spatio-temporal graph convolutional neural network for remaining useful life estimation of aircraft engines. Aerospace Systems **4**, 29–36 (2021)

22. Wang, N., Feng, P., Ge, Z., Zhou, Y., Zhou, B., Wang, Z.: Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems* (2023)
23. Woo, G., Liu, C., Sahoo, D., Kumar, A., Hoi, S.: CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In: *International Conference on Learning Representations* (2022)
24. Yu, W., Kim, I.Y., Mechefske, C.: An improved similarity-based prognostic algorithm for rul estimation using an rnn autoencoder scheme. *Reliability Engineering & System Safety* **199**, 106926 (2020)
25. Zhang, K., Liu, R.: Lstm-based multi-task method for remaining useful life prediction under corrupted sensor data. *Machines* **11**(3), 341 (2023)
26. Zhang, W., Yang, L., Geng, S., Hong, S.: Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems* (2023)
27. Zheng, X., Chen, X., Schurch, M., Mollaysa, A., Allam, A., Krauthammer, M.: Simts: Rethinking contrastive representation learning for time series forecasting. *arXiv preprint [arXiv:2303.18205](https://arxiv.org/abs/2303.18205)* (2023)
28. Zhu, J., Chen, N., Peng, W.: Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Trans. Industr. Electron.* **66**(4), 3208–3216 (2018)
29. Zhu, Q., Xiong, Q., Yang, Z., Yu, Y.: Rgcnu: Recurrent graph convolutional network with uncertainty estimation for remaining useful life prediction. *IEEE/CAA Journal of Automatica Sinica* **10**(7), 1640–1642 (2023)



CCPL: Cross-Modal Contrastive Protein Learning

Jiangbin Zheng^{1,2}  and Stan Z. Li² 

¹ Zhejiang University, Hangzhou, China

² AI Lab, Westlake University, Hangzhou, China
{zhengjiangbin, Stan.ZQ.Li}@westlake.edu.cn

Abstract. Effective protein representation learning is crucial for predicting protein functions. Traditional methods often pretrain protein language models on large, unlabeled amino acid sequences, followed by finetuning on labeled data. While effective, these methods underutilize the potential of protein structures, which are vital for function determination. Common structural representation techniques rely heavily on annotated data, limiting their generalizability. Moreover, structural pretraining methods, similar to natural language pretraining, can distort actual protein structures. In this work, we introduce a novel unsupervised protein structure representation pretraining method, cross-modal contrastive protein learning (CCPL). CCPL leverages a robust protein language model and uses unsupervised contrastive alignment to enhance structure learning, incorporating self-supervised structural constraints to maintain intrinsic structural information. We evaluated our model across various benchmarks, demonstrating the framework's superiority.

Keywords: Protein Representation · Pretrained Language Model · Structure-Sequence Pairing · Contrastive Learning · Unsupervised Learning

1 Introduction

Learning effective protein representations is crucial for various biological tasks. In recent years, deep protein representation learning has revolutionized the field, notably in protein structure prediction, represented by AlphaFold2 [15], and protein design, exemplified by [10] and ProteinMPNN [2]. With the advent of low-cost sequencing technologies, a vast number of new protein sequences have been discovered. Current methods typically pretrain protein language models on large, unlabeled amino acid sequences [18, 24] and then finetuned on downstream tasks using limited labeled data. While sequence-based methods are effective, they often fail to explicitly capture and utilize existing protein structural information, which is vital for protein functions.

Given the high cost and time-consuming nature of annotating new protein functions, there is a pressing need for accurate and efficient function annotation methods to bridge the existing sequence-function gap. Since the functions are

governed by folded structures, some data-driven approaches rely on learning structural representations of proteins, which are then applied to various tasks such as protein design, and function prediction and classification. Therefore, an effective structural encoder is essential. To better leverage structural information, several structure-based protein encoders have been proposed [6,9]. However, due to the scarcity of protein structures, these encoders are often designed for specific tasks, and their generalizability to other tasks remains unclear.

Protein structure encoders face two main challenges: 1) Data scarcity. The number of reported protein structures is significantly lower than datasets in other machine learning fields due to the challenges of experimental protein structure determination. For example, the Protein Data Bank (PDB) contains 182K experimentally determined structures, whereas Pfam has 47 million protein sequences [19] and ImageNet contains 10 million annotated images. 2) Representation difficulty. Unlike sequences, traditional self-supervised language pretraining methods, such as masked language modeling, are not feasible for learning structural representations. Introducing noise or perturbations into structural data can lead to unstable or chemically incorrect structures, making augmented data unreliable. Therefore, the ability to pretrain on known protein structures has not been widely applied to protein property prediction.

By rethinking the protein representations, we observe that the success of sequence-based models is due to large-scale data and the guidance provided by self-supervised signals. Considering there is a natural pairing relationship between structure and sequence, establishing this relationship can help guide structural learning without compromising the protein structure itself. Based on this observation, we pose the question: *Can we augment protein structure model training supervised by robust pretrained protein language models?*

Inspired by advances in cross-modal pretraining (e.g., CLIP [22], Context-to-Vector [30]), we introduce a novel cross-modal contrastive protein learning (CCPL). This method calculates contrastive loss between two independently pretrained encoders, maximizing the similarity between paired protein structures and sequences while minimizing it for non-paired ones [26,28,29,32]. Compared to supervised learning methods, our contrastive learning approach has several advantages. First, finding the matching relationship between protein structures and sequences is natural. Second, our designed contrastive loss reduces dependence on explicit functional annotations, facilitating the use of large-scale unlabeled data. We reframe structural representation training as an information retrieval task, where the protein structure is the query, and the goal is to retrieve the sequence with the highest binding probability to the target protein structure from a pool of candidates. To strengthen structural representation constraints, we also propose a self-supervised contact map constraint based on intermediate features from the structural encoder.

To evaluate our proposed framework, we conducted benchmark tests. Due to the lack of established evaluation strategies for this novel pretraining paradigm, we designed a series of evaluation experiments, including internal tasks (e.g., contact map prediction and distribution alignment quality assessment) to demonstrate internal contrastive alignment ability and refinement performance, and

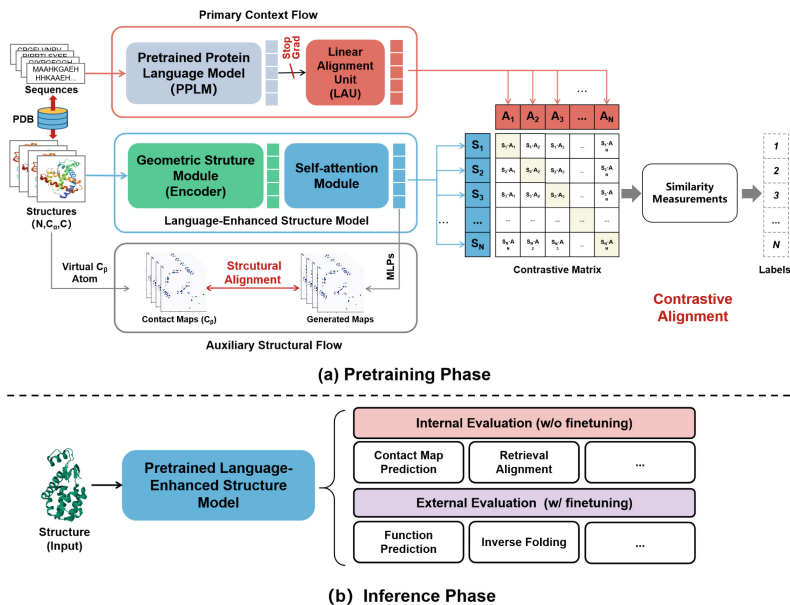


Fig. 1. (a) The proposed cross-modal contrastive learning framework utilizes a pre-trained protein language model to guide the training of the protein structure model through contrastive alignment loss. To reinforce information constraints on the structure, we introduce a self-supervised contact map prediction. (b) The internal and external evaluation tasks for our trained structure model during inference phase.

external/downstream tasks (e.g., protein design and function prediction) to demonstrate generalization capability. Our experimental results validate the effectiveness of the CCPL framework, highlighting its robustness in pretraining performance and exceptional downstream task performance.

Our contributions can be summarized as follows:

- We propose an cross-modal protein representation framework, establishing a novel deep alignment relationship between sequences and structures. For the first time, we pretrain protein structural models under the guidance of rich prior language knowledge from pretrained protein sequence models.
- We introduce a comprehensive evaluation system to assess the pretrained structural models, providing benchmarks for the protein research community.
- Our proposed protein structural model pretraining demonstrates competitive performance across various evaluation tasks.

2 Proposed CCPL Framework

2.1 Problem Statements

Viewing CCPL as a supervised pseudo-dense retrieval task, we treat the protein structure as a query and retrieve the most relevant sequence from a given

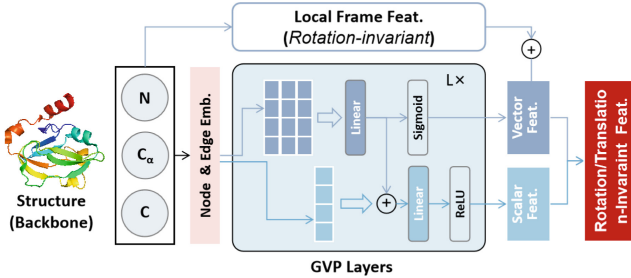


Fig. 2. Schematic diagram of the GVP module: Protein backbone atoms (C, C_α , and N) form the basis for generating graphs with node and edge features based on k-nearest neighbor relationships. These graphs are fed into the vector and scalar channels of the GVP module to produce vector and scalar features. These primary features are then enhanced with additional spatial features, including rotation frame, sidechain, orientation, and dihedral characteristics, to create comprehensive spatial structure features.

dataset. The overall framework, illustrated in Figure 1, involves training two separate encoders to learn representations for protein structures and sequences, respectively. The similarity between each protein structure-sequence pair is then calculated, and a contrastive learning objective is used to distinguish between positive and negative pairs. Formally, given a protein structure p and its paired protein sequence m , the objective of CCPL representation training is to maximize the probability of the sequence that naturally pairs with the structure. This selection process is guided by a scoring function $\delta(\cdot)$, which evaluates the pairing probability between the protein p and the candidate sequence m .

2.2 Protein Structure and Sequence Encoders

Equivariant Protein Structure Encoder. For downstream protein structure tasks, it is essential that the predicted sequences remain unaffected by the reference frame of the structural coordinates. This means that the model’s output distribution should be invariant under any rotation or translation applied to the input coordinates H . Equivariant network models, such as geometric vector perceptron (GVP)-based models, are commonly employed in protein 3D structure modeling to meet this requirement. Models like GVP-GNN [14] and GVP-Transformer [10] have demonstrated impressive performance in protein design tasks, highlighting the crucial role of the GVP module in representing structural features while maintaining equivariant properties. Thus, we use GVP as our equivariant structure module. For any transformation $T \in E(3)$, the GVP module φ is considered $E(3)$ -equivariant since $\varphi(T \cdot H) = T \cdot \varphi(H)$, and $E(3)$ -invariant since $\varphi(T \cdot H) = \varphi(H)$. Moreover, the simplicity and lightweight nature of GVP components make the GVP architecture well-suited for our structure module. A schematic diagram of the GVP module is illustrated in Figure 2.

Pretrained Protein Language Encoder. Pretrained protein language models have become integral to deep protein downstream tasks, having been trained on extensive datasets. These sequence models inherently encapsulate rich structural information gleaned from sequence data, enabling them to effectively guide structure model learning. Among the available models, we choose ESM-2 [18] as our primary teacher model due to its exceptional performance.

Formulation. Formally, we denote the protein structure encoder as g_ϕ with parameters ϕ and the sequence encoder as f_θ with parameters θ . The representations of the protein structure vector x^p and the corresponding sequence vector y^m are then defined as $g_\phi(x^p)$ and $f_\theta(y^m)$, respectively.

2.3 Training Objectives

Contrastive Alignment Objective. To enable the contrastive learning process, we first need to measure the similarity between each protein structure and sequence pair. Drawing on previous research, we can utilize either dot product or cosine similarity for this purpose. When using the dot product, the similarity score for pairs (x_p^i, y_m^j) , where $i, j \in [1, N]$, is defined as:

$$\delta(x_p^i, x_m^j) = g_\phi(x_p^i)^T \cdot f_\theta(y_m^j), \quad (1)$$

where, by normalizing these scores, we then obtain cosine similarity. Since our protein dataset includes only positive pairs of binding protein structures and sequences, we need to create negative pairs for contrastive learning. We implement a batch-wise sampling strategy inspired by CLIP. For a given batch of paired data $\{(x_b^p, y_b^m)\}_{b=1}^B$ with batch size B , we extract a list of protein structures $\{x_b^p\}_{b=1}^B$ and a corresponding list of sequences $\{y_b^m\}_{b=1}^B$. By combining these lists, we generate B^2 pairs (x_i^p, y_j^m) where $i, j \in [1, B]$. Pairs where $i = j$ are positive, while pairs where $i \neq j$ are negative. This approach relies on a fundamental assumption: if a protein and sequence pair is known to bind, it is likely that this protein does not bind with other sequences and vice versa. This assumption is validated by the distinct distribution patterns of positive and negative pairs.

We formalize this with two loss functions: the structure-to-sequence loss \mathcal{L}^p and the sequence-to-structure loss \mathcal{L}^m . The structure-to-sequence loss quantifies the likelihood of ranking the correct binding sequence higher than other sequences for a given protein structure x_p^b :

$$\mathcal{L}_b^p(x_p^b, \{y_i^m\}_{i=1}^B) = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\delta(x_p^b, y_i^m))}{\sum_{j=1}^B \exp(\delta(x_p^b, y_j^m))} \quad (2)$$

Conversely, the sequence-to-structure loss measures the likelihood of correctly ranking the binding target for a given molecule y_m^b :

$$\mathcal{L}_b^m(y_m^b, \{x_i^p\}_{i=1}^B) = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\delta(y_m^b, x_i^p))}{\sum_{j=1}^B \exp(\delta(y_m^b, x_j^p))} \quad (3)$$

Therefore, the final contrastive alignment loss for a mini-batch is the average of the two-direction losses $\mathcal{L}_{\text{align}}$ as:

$$\mathcal{L}_{\text{align}} = \frac{1}{2} \sum_{b=1}^B (\mathcal{L}_b^p + \mathcal{L}_b^m). \quad (4)$$

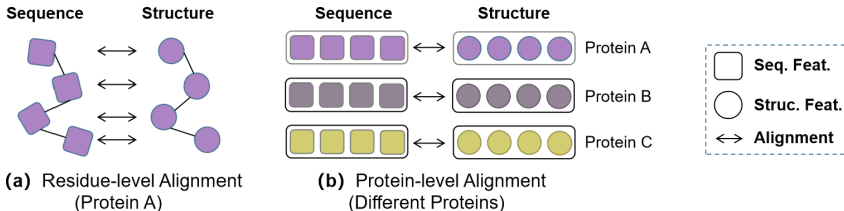


Fig. 3. Various alignment levels. (a) Residue-level alignment entails comparing each pair of structure-sequence features residue by residue. (b) Protein-level alignment involves comparing each pair of structure-sequence features protein by protein. The features of each protein are amalgamated from all the residue features contained within it. Identical colors indicate a sequence-structure pair originating from the same protein.

Contrastive Alignment Level. As shown in Figure 3, we propose to compute the alignment loss at both the residue and protein levels, respectively. For residue-level alignment, we compare the encoded sequence and structure features residue by residue. In contrast, for protein-level alignment, we compare the encoded features at a coarser, fine-grained level. Intuitively, alignment at different fine-grained levels results in different capabilities. Finer-grained comparisons, such as residue-level alignment, necessitate more complex computations but may yield better performance. Conversely, coarse-grained comparison alignments, such as protein-level alignment, may exhibit inferior performance but are worth considering due to their lighter computational load. Refer to our experimental section for a detailed analysis. Regardless of the level used for calculation, the samples in the mini-batch will be randomly shuffled, and each pairing will be different, thereby implicitly serving as a form of data augmentation.

Structural Reconstruction Constraint. As previously mentioned, the GVP layers encode the core structural features using N, C_α , and O atoms (These three types of atoms are called backbone atoms). To further enhance the structural constraints, we introduce a self-supervised structure reconstruction task. While predicting the coordinates of virtual C_β atoms directly from the structural features of N, C_α , and O atoms would be the most straightforward approach, it proves challenging to achieve accurate predictions in practice. Therefore, we simplify this task by transforming the coordinate prediction into a contact map prediction, illustrated in the Figure 4. Specifically, we utilize the intermediate attention maps generated by the self-attention blocks depicted in Figure 1(a)

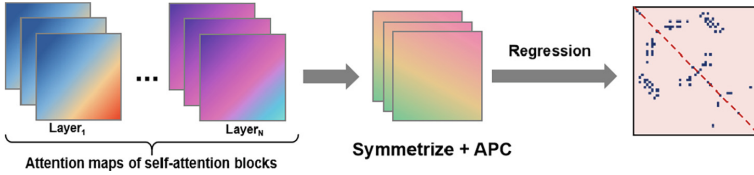


Fig. 4. Pipeline for reconstructing the contact map based on C_β atoms with length L : First, attention maps are extracted from each layer of the self-attention blocks. These maps undergo symmetrization and average product correction (APC) along the amino acid dimensions to produce an $L \times L$ coupling matrix. This matrix forms the basis for the final contact map predictions, which are refined using a regression layer.

to predict the contact maps related to C_β [24]. In addition to serving as a constraint during training to preserve structural information, the generation of contact maps can also serve as a metric for internally evaluating the performance of the pretrained model.

Virtual C_β Atom Generation. The virtual C_β atoms are derived coordinates that may not physically exist in every residue. However, their positions can be inferred from the spatial relationships among protein backbone atoms: N, C_α , and O atoms. The positions are calculated as follows:

$$\begin{cases} \delta_0 = \varrho(C_\alpha) - \varrho(N), \\ \delta_1 = \varrho(C) - \varrho(C_\alpha), \\ \delta_2 = \delta_0 \otimes \delta_1, \\ \varrho(C_\beta) = \epsilon_0 \cdot \delta_0 + \epsilon_1 \cdot \delta_1 + \epsilon_2 \cdot \delta_2 + \varrho(C_\alpha), \end{cases} \quad (5)$$

where \otimes denotes the cross product of vectors and $\varrho(*)$ denotes the coordinates of the corresponding atom types. And $\epsilon_0 = 0.5680$, $\epsilon_1 = -0.5407$ and $\epsilon_2 = -0.5827$ are constant coefficients.

Self-supervised Contact Map Prediction. The pipeline diagram of the contact map predictor is depicted in Figure 4. Self-attention maps extracted from the self-attention blocks undergo symmetrization and average product correction (APC) to generate the final contact maps. The self-supervised structural loss is formulated to constrain the structural feature representation using cross-entropy:

$$\begin{aligned} \mathcal{L}_{contact} &= \text{CrossEntropy}(\text{Sigmoid}(M_{pred}), M_{ref}) \\ &= - \sum_{i=1} M_{ref}^{(i)} \cdot \log(M_{pred}^{(i)}), \end{aligned} \quad (6)$$

where M_{ref} denotes the ground-truth contact maps (binary values 0 or 1), and M_{pred} denotes the predicted values distributed between 0 and 1.

The pretraining objective involves jointly optimizing the contrastive alignment loss \mathcal{L}_{align} and the contact map loss $\mathcal{L}_{contact}$, with weighted values λ_1 and λ_2 , as expressed by::

$$\mathcal{L}_{pretrain} = \lambda_1 \cdot \mathcal{L}_{align} + \lambda_2 \cdot \mathcal{L}_{contact}. \quad (7)$$

2.4 Inference Phase

As depicted in Figure 1(b), the architecture of the inference phase utilizes only the pretrained language-enhanced structure model, rendering other flows unnecessary during this stage. Evaluating a pretrained protein structure model within a novel training framework poses significant challenges. To address these challenges, we have developed a comprehensive evaluation system that includes multiple validation tasks, showcasing the model’s representation learning, alignment capability, and generalization ability. Based on the necessity for fine-tuning, we categorize the validation tasks into internal and external/downstream tasks.

External Evaluation Tasks. External tasks require fine-tuning and are focused on downstream applications. We introduce the protein sequence design task (also known as protein inverse folding), which involves predicting protein sequences based on corresponding protein backbone atomic coordinates. Key metrics for this task include perplexity and accuracy of sequence recovery. Additionally, we incorporate protein functional recognition tasks to validate the robust representation capabilities acquired during training.

Internal Evaluation Tasks. Internal tasks do not require fine-tuning and include contact map prediction and self-similarity evaluation. The contact map prediction task uses the top-L long-range precision (P@L) metric to evaluate the quality of the predicted contact maps. The self-similarity evaluation task assesses the alignment between the language model and the structure model using accuracy and KL divergence metrics.

3 Experiments

Table 1. F_{max} of gene ontology term prediction and enzyme commission prediction.

Input	Methods	Gene Ontology			Enzyme Commission
		BP	MF	CC	
1D	CNN	0.244	0.354	0.287	0.545
	ResNet	0.280	0.405	0.304	0.605
	LSTM	0.225	0.321	0.283	0.425
	Transformer	0.264	0.211	0.405	0.238
1D	GCN	0.252	0.195	0.329	0.320
	GAT	0.284	0.317	0.385	0.368
	3D CNN	0.240	0.147	0.305	0.077
3D+1D	GraphQA	0.308	0.329	0.413	0.509
	GVP [13]	0.326	0.426	0.420	0.489
	IEConv (residue level) [8]	0.421	0.624	0.431	-
	GearNet [25]	0.356	0.503	0.414	0.730
	CDCConv [3]	0.453	0.654	0.479	0.820
3D	Ours	0.459	0.663	0.491	0.828

3.1 Settings

Pretraining Datasets. We utilize a larger-scale protein dataset, PDB, comprising sequence-structure pairs, as the pretraining dataset. To prevent label

leakage, we exclude all existing data also appearing in evaluation datasets. To augment the datasets, we use the AlphaFoldDB for pretraining. This database contains the protein structures predicted by the AlphaFold2 model.

Evaluation Datasets. The CATH dataset is widely employed, featuring training, validation, and test splits consisting of 18204, 608, and 1120 protein data samples. In downstream protein design tasks, the training set is utilized for fine-tuning, and the test set is used for evaluation. We also report results on Ts50 & Ts500 [16]. Furthermore, the trRosetta set is utilized for contact map prediction, comprising around 15000 instances. The CASP14 set, renowned for AlphaFold2, though modest in number, closely resembles the practical environment of blind tests and competitions. To summarize, the trRosetta set, CATH testing set, Ts50/Ts500, and CASP14 set are all utilized for internal evaluation.

Implementation Details. During pretraining, AdamW optimizer with a batch size of 8 and an initial learning rate of 1e-3 is used. We employ the ESM-2 base version as the default teacher protein language model, with fixed parameters in the training pipeline. The GVP module comprises 4 layers with a dropout of 0.1, top-k neighbors of 30, a node hidden dimension of scalar features of 1024, and a node hidden dimension of vector features of 256. The self-attention block following the GVP includes 4 self-attention layers, 8 attention heads, an embedded dimension of 512, and an attention dropout of 0.1. It’s noteworthy that we pretrain the model separately for residue-level alignment and protein-level alignment. 2 NVIDIA GPU A100 80GB were used.

3.2 Evaluation on Protein Function Prediction Tasks

Table 2. Comparison among our protein design models (#2) and baselines (#1). The best results are **bolded**, followed by underlined. Design_p: Protein-level pretrained model; Design_r: Residue-level pretrained model.

#	Models	Perplexity			Recovery (%)		
		CATH	Ts50	Ts500	CATH	Ts50	Ts500
1	Natural frequencies [10]	17.97	-	-	9.5	-	-
	SPIN2 [21]	-	-	-	-	33.6	36.6
	Structured Transformer [12]	6.85	5.60	5.16	36.4	42.40	44.66
	Structured GNN [14]	6.55	5.40	4.98	37.3	43.89	45.69
	GVP-Transformer [10]	6.44	-	-	38.3	-	-
	AlphaDesign [4]	6.30	5.25	4.93	41.31	48.36	49.23
	GVP-GNN-large [14]	6.17	-	-	39.2	-	-
	GVP-GNN [14]	5.29	4.71	4.20	40.2	44.14	49.14
	ProteinMPNN [2]	4.61	3.93	3.53	45.96	54.43	58.08
	2	Design (<i>w/o</i> Pretraining)	6.27	5.05	4.87	39.62	49.21
Design _p (<i>w/</i> Pretraining)		<u>4.51</u>	<u>3.82</u>	<u>3.35</u>	<u>50.1</u>	<u>55.7</u>	<u>59.5</u>
Design _r (<i>w/</i> Pretraining)		4.48	3.76	3.28	50.8	55.8	60.3

Our method shows superior performance in gene ontology (GO) term prediction and enzyme commission (EC) number prediction tasks, surpassing existing 1D-only, 3D-only, and (3+1)D approaches. We used the F_{max} accuracy metric for

evaluation. These findings, detailed in Table 1, highlight the effectiveness of our cross-modal contrastive learning approach.

The IEConv method, featuring both atom-level (Hermosilla et al., 2021) and amino-acid-level (Hermosilla & Ropinski, 2022) variants, served as a key benchmark in our comparisons. The atom-level variant, denoted as "3D+Topo," utilizes 3D coordinates and the topological structure of bonds between atoms. Our approach outperformed this and other existing methods significantly. Following the work of Hermosilla et al. (2021), Hermosilla & Ropinski (2022), and Zhang et al. (2022), we assessed our method across three GO term prediction sub-tasks: biological process (BP), molecular function (MF), and cellular component (CC). Both GO term and EC number predictions are framed as multi-label classification tasks. In essence, our approach integrates 1D and 3D data using cross-modal contrastive learning, yielding robust representations and achieving higher accuracy in GO and EC prediction tasks.

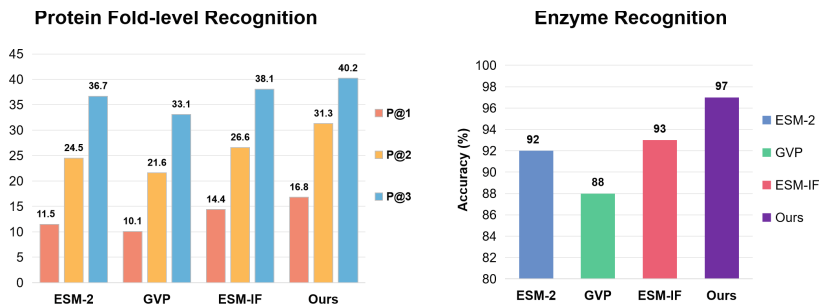


Fig. 5. Protein functional prediction tasks. (a) Comparing the fold-level predictions. P@k denotes the top-k precision. (b) Performances for enzyme recognition task.

3.3 Evaluation on Protein Inverse Folding Tasks

Computational protein design, also known as *protein inverse folding*, aims to deduce amino acid sequences given the corresponding atomic coordinates of protein backbones. The protein design model serves as a key component for downstream evaluation. It directly leverages the pretrained structure model as the backbone, coupled with a non-autoregressive decoder featuring linear MLPs. The CATH testing set and Ts50/Ts500 are employed to evaluate the primary results. Table 2 presents a selection of several prominent baselines across different types. Notably, ProteinMPNN [2] and GVP-Transformer [10] exhibit advanced performance, particularly excelling in sequence recovery and perplexity. The lightweight GVP-GNN [14] also demonstrates competitiveness, showcasing relatively strong performance and speed. Our protein design models outperform the baselines. Notably, the inclusion of a non-autoregressive decoder in our module contributes to faster sampling speeds. Furthermore, regarding different levels (Design_p vs. Design_r), while there is no significant disparity between the

residue-level and protein-level pretrained modules in downstream tasks, Design_r slightly outperforms Design_p overall. This observation suggests that any small gaps between the two levels of pretrained models are further diminished during the fine-tuning process. Additionally, within Group 2 of Table 2, we compared the pretrained model with a non-trained model serving as the backbone. Notably, the pretrained model significantly enhances performance in terms of perplexity and recovery, underscoring the stronger generalization ability of the pretrained structure model for downstream tasks.

3.4 Evaluation on Protein Fold-level Classification

Furthermore, we undertake predictions on challenging fold types based on the Fold dataset [7], which essentially involves a less data-intensive multi-task enzyme function prediction. We gather approximately 700 enzymes with experimentally determined structure-sequence pairs across 10 folds from RCSB for multi-class functional prediction tasks. This involves precisely predicting the fold level to which an enzyme belongs. ‘Fold’ here refers to the 3D arrangement of secondary structural elements (such as alpha helices and beta sheets) that characterize a particular protein or group of proteins. Proteins with similar folds typically share significant structural similarities, even if their sequences and functions differ. Fold classification can offer insights into evolutionary relationships among proteins. The objective of this setup is to predict, within enzymes with mixed folds, the specific fold to which an enzyme belongs. As illustrated in Figure 5(a), we choose ESM-IF and GVP as baseline structure-to-sequence models, as they also utilize N, C_α , and C as standard inputs. These comparisons validate the superior structural representation capabilities by a large margin (Ours: 16.8%P@1, 31.3%P@2, 40.2%P@3). We employ ESM-2 as a language modality input for comparison to confirm its sequence-only representation capabilities. Although ESM-2 (11.5%P@1, 24.5%P@2, 36.5%P@3) slightly outperforms the earlier GVP, it significantly lags behind the performance of ours, demonstrating enhanced generalization ability due to the incorporation of context information.

3.5 Evaluation on Functional Enzyme Recognition

To explore enzyme recognition through binary classification, we leverage a novel dataset that extends the Fold dataset. Each fold in this dataset is meticulously crafted to include an equal number of enzymes, balanced with non-enzyme negative samples. Given the binary nature of the classification task, positive and negative samples are aggregated across folds and then randomly divided, with 80% allocated to the training set and 20% to the evaluation set. Our experimental findings, as illustrated in Figure 5(b), reveal that our model, augmented with prior language enhancements, achieves the highest performance, boasting an average accuracy of 97%. Among the baseline models, ESM-IF and ESM2 yield comparable accuracies of 93% and 92%, respectively, despite operating on different modalities (sequence vs. structure). Notably, our model outperforms ESM-IF, a benchmark recognized for its exceptional representation and similarity in modalities, indicating superior generalization capabilities.

3.6 Zero-shot Learning for Protein Fitness Prediction

Additionally, we introduce a zero-shot learning approach for fitness prediction, which enables the direct validation of the model’s representational stability in a non-parametric manner. To ensure a thorough and unbiased comparison, we selected prominent protein language models and inverse folding models as benchmarks, as depicted in Figure 6. All baseline models utilize zero-shot learning for evaluating mutation effects through the ProteinGym dataset [20], which encompasses millions of mutations. Our proposed method yields an average ρ of 43.0%, consistently achieving the top matching rank. This suggests that leveraging prior language knowledge significantly contributes to enhanced overall performance in mutation prediction. In contrast, the optimal baseline model (ESM-IF) scores 42.2%. The superiority in performance of our model can be attributed to its dual advantage, encompassing both contextual and structural transfer learning within the proposed training paradigm.

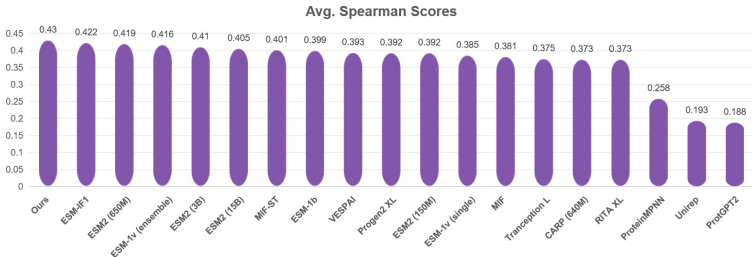


Fig. 6. Spearman’s rank correlation on ProteinGym set.

4 Ablation Studies and Analysis

Table 3. Internal evaluation across test sets. Group 1 showcases the contact map predictions, measured by P@* scores. Group 2 focuses on retrieval alignment evaluations, quantified by alignment accuracy and KL distance. The acc_1 and acc_2 metrics denote the accuracy of structure-to-sequence and sequence-to-structure alignment.

Group	Level	CATH Test Set			trRosetta Set			Ts500 Set			Ts50 Set			CASP14 Set		
		P@L	P@L2	P@L5	P@L	P@L2	P@L5	P@L	P@L2	P@L5	P@L	P@L2	P@L5	P@L	P@L2	P@L5
1	Residue	78.05	90.97	96.12	87.69	94.94	96.89	90.31	96.67	98.10	91.36	98.66	100.00	74.9	91.49	95.34
	Protein	72.21	88.03	98.31	81.30	92.42	96.18	83.78	94.14	97.44	85.18	96.32	99.53	68.24	87.23	93.96
2	Residue	acc_1	acc_2	KL	acc_1	acc_2	KL	acc_1	acc_2	KL	acc_1	acc_2	KL	acc_1	acc_2	KL
	Protein	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00

4.1 Internal Contact Map Generation

Contact map predictions serve as a significant indicator of structural representation capabilities. Due to our pretraining mechanism, the contact map predictor

can generate accurate contact maps directly without the need for fine-tuning. Group 1 of Table 3 showcases the contact map prediction scores, evaluated across the CATH, trRosetta, Ts50/Ts500, and CASP14 test sets. Both the residue-level and protein-level pretrained models demonstrate high P@L accuracy in predicting contact maps across all datasets, indicating that the pretrained structure module has acquired rich structural representations. Notably, the residue-level evaluation exhibits superior performance within Group 1, likely attributable to its finer granularity. Expanding on this analysis, the ability of our pretrained models to generate precise contact maps across diverse datasets underscores their robustness and generalization potential in capturing intricate structural details. Such proficiency in contact map prediction signifies the effectiveness of our pretraining approach in enhancing structural representation learning.

4.2 Retrieval Alignment Evaluation

In the proposed pretraining framework, we operate under the strong assumption that protein language models and protein structure models are equally proficient in representing features, albeit through different modalities. Hence, we advocate for quantifying the sequence-structure retrieval power to gauge the alignment prowess of the pretrained model. Reflecting on the contrastive alignment loss employed during pretraining, it becomes evident that the loss encompasses both structure-to-sequence and sequence-to-structure alignment calculations. The intermediate state score computations offer a direct means to evaluate the multi-modality alignment level. The retrieval alignment evaluation on the CATH, trRosetta, Ts50/Ts500, and CASP14 test sets is presented in Group 2 of Table 3. Additionally, we provide residue-level and protein-level results for comprehensive analysis. It’s noteworthy that the protein-level pretrained model exhibits a higher ease in aligning sequences and structures, evident through higher accuracy and lower KL distance, which aligns well with our intuition. Overall, the pretrained structure module demonstrates a robust alignment level, underscoring the effectiveness of the proposed framework.

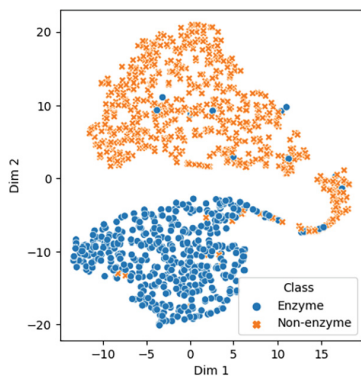


Fig. 7. Visualization using t-SNE to demonstrate enzyme recognition.

4.3 Classification Visualization

To visually represent the efficacy of enzyme classification, we advocate the utilization of t-distributed Stochastic Neighbor Embedding (t-SNE) as a means to illustrate the clustering patterns of enzymes, as depicted in Figure 7. Employing this visualization technique provides insights into the distribution of enzyme data points in a lower-dimensional space. In the context of enzyme recognition, our proposed structural model manifests discernible classification boundaries within the t-SNE plot. This observation underscores the robustness of our model’s feature representation capabilities. The distinct clustering of enzymes reaffirms the model’s ability to delineate between enzyme and non-enzyme samples effectively, further validating suitability for the task of functional recognition.

5 Related Work

Protein Language Models. Protein language modeling has emerged as a promising avenue for unsupervised learning of protein primary structures [11, 27, 31]. Models such as UniRep [1] utilize LSTM or its variants to capture sequence representations and long-range dependencies. TAPE [23] benchmarks a range of protein models across various tasks, affirming the effectiveness of self-supervised pretraining methods. Many efforts have focused on enhancing model scale and architecture to capture richer protein semantics. For instance, ESM-1b employs a Transformer architecture and masked language modeling strategy to learn robust representations from a large-scale dataset. Subsequently, ESM-2 [17] extends ESM-1b with larger-scale parameters (15 billion), achieving superior results compared to smaller ESM models.

Protein Structure Models. Given that a protein’s function is often determined by its structure. The advancements in protein structure prediction methods have led to initiatives like AlphaFoldDB, providing over 200 million protein structure predictions to accelerate scientific research. Building upon this progress, various protein structure models have emerged, aiming to encode spatial information using convolutional neural networks (CNNs) or graph neural networks (GNNs). Among these methods, IEConv [8] introduces a convolution operator to capture all relevant structural levels of a protein. GearNet [25] encodes the spatial information by adding different types of sequential or structural edges and then performed relational message passing on protein residue graphs. GVP-GNN [13] designed the geometric vector perceptrons (GVP) to learn both scalar and vector features in an equivariant and invariant manner, while [5] adopts SE(3)-invariant features as model inputs and reconstruct gradients over 3D coordinates to avoid the complexity of SE(3)-equivariant models.

6 Conclusions and Limitations

We propose leveraging pretrained protein language model to train protein structure models using cross-modal contrastive learning. Our approach demonstrates superior performances in various evaluation tasks. However, challenges remain, including the scope of language model transfer, data efficiency, generalization, computational resources, and evaluation metrics. Addressing these limitations will be crucial for advancing the utility of pretrained protein language models in protein structure prediction and related applications.

Acknowledgments. This work was supported by National Science and Technology Major Project (No. 2022ZD0115101), National Natural Science Foundation of China Project (No. U21A20427), Project (No. WU2022A009) from the Center of Synthetic Biology and Integrated Bioengineering of Westlake University and Integrated Bioengineering of Westlake University and Project (No. WU2023C019) from the Westlake University Industries of the Future Research Funding.

References

1. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M.: Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* (2019)
2. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I., Courbet, A., de Haas, R.J., Bethel, N., et al.: Robust deep learning-based protein sequence design using proteinmpnn. *Science* **378**(6615), 49–56 (2022)
3. Fan, H., Wang, Z., Yang, Y., Kankanhalli, M.: Continuous-discrete convolution for geometry-sequence modeling in proteins. In: *The Eleventh International Conference on Learning Representations* (2022)
4. Gao, Z., Tan, C., Li, S., et al.: Alphadesign: A graph protein design method and benchmark on alphafolddb. *arXiv preprint [arXiv:2202.01079](https://arxiv.org/abs/2202.01079)* (2022)
5. Guo, Y., Wu, J., Ma, H., Huang, J.: Self-supervised pre-training for protein embeddings using tertiary structures (2022)
6. Hermosilla, P., Ropinski, T.: Contrastive representation learning for 3d protein structures. *arXiv preprint [arXiv:2205.15675](https://arxiv.org/abs/2205.15675)* (2022)
7. Hermosilla, P., Schäfer, M., Lang, M., Fackelmann, G., Vázquez, P.P., Kozlíková, B., Krone, M., Ritschel, T., Ropinski, T.: Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv preprint [arXiv:2007.06252](https://arxiv.org/abs/2007.06252)* (2020)
8. Hermosilla, P., Schäfer, M., Lang, M., Fackelmann, G., Vázquez, P.P., Kozlíková, B., Krone, M., Ritschel, T., Ropinski, T.: Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *Learning* (2020)
9. Hermosilla, P., Schfer, M., Lang, M., Fackelmann, G., Vázquez, P.P., Kozlikova, B., Krone, M., Ritschel, T., Ropinski, T.: Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures (2021)
10. Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., Rives, A.: Learning inverse folding from millions of predicted structures. *bioRxiv* (2022)
11. Hu, B., Tan, C., Xia, J., Zheng, J., Huang, Y., Wu, L., Liu, Y., Xu, Y., Li, S.Z.: Learning complete protein representation by deep coupling of sequence and structure. *bioRxiv pp. 2023–07* (2023)

12. Ingraham, J., Garg, V., Barzilay, R., Jaakkola, T.: Generative models for graph-based protein design. *Advances in neural information processing systems* **32** (2019)
13. Jing, B., Eismann, S., Suriana, P., Townshend, R.J.L., Dror, R.O.: Learning from protein structure with geometric vector perceptrons. *Learning* (2020)
14. Jing, B., Eismann, S., Suriana, P., Townshend, R.J., Dror, R.: Learning from protein structure with geometric vector perceptrons. [arXiv:2009.01411](https://arxiv.org/abs/2009.01411) (2020)
15. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. *Nature* **596**(7873), 583–589 (2021)
16. Li, Z., Yang, Y., Faraggi, E., Zhan, J., Zhou, Y.: Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics* **82**(10), 2565–2573 (2014)
17. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Costa, A.D.S., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Language models of protein sequences at the scale of evolution enable accurate structure prediction (2022)
18. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al.: Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* (2022)
19. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., Tosatto, S.C., Paladin, L., Raj, S., Richardson, L.J., et al.: Pfam: The protein families database in 2021. *Nucleic acids research* **49**(D1) (2021)
20. Notin, P., Kollasch, A.W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Weitzman, R., Frazer, J., et al.: Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv* pp. 2023–12 (2023)
21. O’Connell, J., Li, Z., Hanson, J., Heffernan, R., Lyons, J., Paliwal, K., Dehzangi, A., Yang, Y., Zhou, Y.: Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics* **86**(6), 629–633 (2018)
22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. [arXiv:2204.06125](https://arxiv.org/abs/2204.06125) (2022)
23. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., Song, Y.S.: Evaluating protein transfer learning with tape. *bioRxiv* (2019)
24. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., Rives, A.: Transformer protein language models are unsupervised structure learners. *Biorxiv* (2020)
25. Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., Tang, J.: Protein representation learning by geometric structure pretraining (2022)
26. Zheng, J., Chen, Y., Wu, C., Shi, X., Kamal, S.M.: Enhancing neural sign language translation by highlighting the facial expression information. *Neurocomputing* **464**, 462–472 (2021)
27. Zheng, J., Li, S., Huang, Y., Gao, Z., Tan, C., Hu, B., Xia, J., Wang, G., Li, S.Z.: Mmdesign: Multi-modality transfer learning for generative protein design. *arXiv preprint* [arXiv:2312.06297](https://arxiv.org/abs/2312.06297) (2023)
28. Zheng, J., Li, S., Tan, C., Wu, C., Chen, Y., Li, S.Z.: Leveraging graph-based cross-modal information fusion for neural sign language translation. *arXiv preprint* [arXiv:2211.00526](https://arxiv.org/abs/2211.00526) (2022)
29. Zheng, J., Wang, Y., Tan, C., Li, S., Wang, G., Xia, J., Chen, Y., Li, S.Z.: Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. *arXiv preprint* [arXiv:2303.05725](https://arxiv.org/abs/2303.05725) (2023)

30. Zheng, J., Wang, Y., Wang, G., Xia, J., Huang, Y., Zhao, G., Zhang, Y., Li, S.Z.: Using context-to-vector with graph retrofitting to improve word embeddings. arXiv preprint [arXiv:2210.16848](https://arxiv.org/abs/2210.16848) (2022)
31. Zheng, J., Zhang, H., Xu, Q., Zeng, A.P., Li, S.Z.: Metaenzyme: Meta pan-enzyme learning for task-adaptive redesign. arXiv preprint [arXiv:2408.10247](https://arxiv.org/abs/2408.10247) (2024)
32. Zheng, J., Zhao, Z., Chen, M., Chen, J., Wu, C., Chen, Y., Shi, X., Tong, Y.: An improved sign language translation model with explainable adaptations for processing long sign sentences. *Computational Intelligence and Neuroscience* **2020** (2020)



Box2Flow: Instance-Based Action Flow Graphs from Videos

Jiatong Li^(✉), Kalliopi Basioti, and Vladimir Pavlovic

Rutgers University, Piscataway, NJ 08854, USA
jiatong.li@rutgers.edu

Abstract. A large amount of procedural videos on the web show how to complete various tasks. These tasks can often be accomplished in different ways and step orderings, with some steps able to be performed simultaneously, while others are constrained to be completed in a specific order. Flow graphs can be used to illustrate the step relationships of a task. Current task-based methods try to learn a single flow graph for all available videos of a specific task. The extracted flow graphs tend to be too abstract, failing to capture detailed step descriptions. In this work, our aim is to learn accurate and rich flow graphs by extracting them from a single video. We propose **Box2Flow**, an instance-based method to predict a step flow graph from a given procedural video. In detail, we extract bounding boxes from videos, predict pairwise edge probabilities between step pairs, and build the flow graph with a spanning tree algorithm. Experiments on MM-ReS and YouCookII show our method can extract flow graphs effectively.

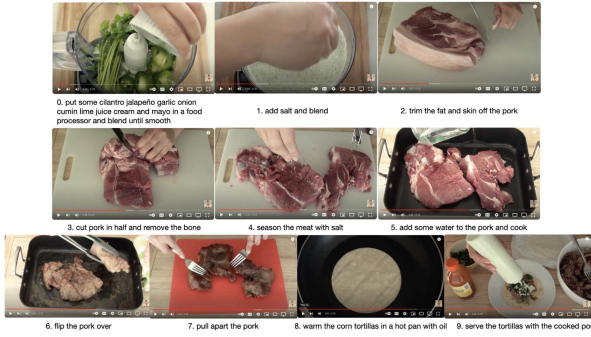
Keywords: Flow Graph · Procedural Videos · Object Detection

1 Introduction

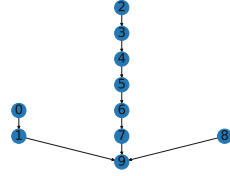
Procedural videos showing how to perform various tasks can be found on video sharing platforms, ranging from adding oil to cars to making cakes. This wealth of data creates an opportunity for computer vision systems [37, 38, 40] to learn a computational representation of those multi-step procedures, which can then be used in various downstream applications ranging from video activity segmentation to general procedure analytics.

However, in real-world procedures seemingly identical tasks are often performed differently by individual users, including, e.g., using different materials or cooking ingredients, different actions, different step orderings, and different number of steps, while also sharing some common procedural elements. This will lead to different procedure workflows depending on each instance of the task as recorded in a video. As shown in Figure 1a, 1c, two recipes, **Carnitas Tacos**

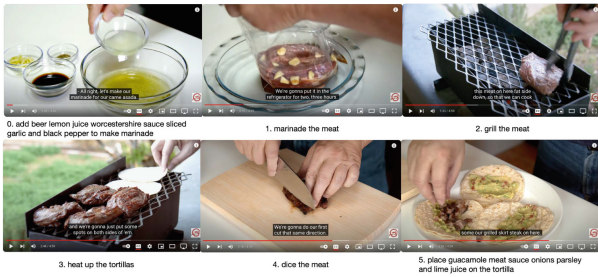
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78398-2_3.



(a) Carnitas Tacos With Cilantro Lime Sauce.



(b) Flow Graph of Carnitas Tacos With Cilantro Lime Sauce.



(c) Carne Asada Tacos.



(d) Flow Graph of Carne Asada Tacos.

Fig. 1. (a), (c): Two recipes for making tacos that differ in ingredients, actions, and number of steps. (b), (d): The corresponding flow graphs of the two recipes.

With **Cilantro Lime Sauce** and **Carne Asada Tacos**, both belong to the same task category, making tacos, but they are very different. The two recipes use different ingredients for flavoring. **Carnitas Tacos With Cilantro Lime Sauce** added the sauce in the last step while **Carne Asada Tacos** marinated the meat. In terms of actions, **Carnitas Tacos With Cilantro Lime Sauce** cooked the pork with water and pulled apart the pork while **Carne Asada Tacos** grilled the tortillas and diced the meat. **Carnitas Tacos With Cilantro Lime Sauce** heated the tortillas after pulling apart the pork while **Carne Asada Tacos** heated the tortillas before dicing the meat. Finally, **Carnitas Tacos With Cilantro Lime Sauce** is annotated with 10 steps while **Carne Asada Tacos** is annotated with 6. As a result, their workflows are also different, as in Figure 1b, 1d. Therefore, in order to comprehend procedural videos, a computer vision system must be able to recognize the various types of steps and their possible sequences. We propose that disassembling each video instance separately into individual steps can lead to a better understanding of the overall task than task-based methods [15, 21, 37] that try to learn task steps simultaneously by processing all available videos of a particular task.

One aspect of understanding the procedure flow in the video is determining the step dependencies. Some earlier steps are prerequisites of later steps. For example, in *Carnitas Tacos With Cilantro Lime Sauce*, the sauce corresponding to steps 0 and 1 must be made, the meat corresponding to steps 2-7 must be cooked, and the tortilla corresponding to step 8 must be heated in order for the taco to be assembled in step 9. Therefore, steps 1, 7, and 8 are prerequisites of step 9, and this relationship is defined as **sequential**. Meanwhile, steps 1, 7, and 8 deal with three parallel components of the taco, and switching their order will not affect the final dish. In other words, the step sequence 0,2,1,3,4,5,6,8,7,9 would also be a valid recipe resulting in the same dish. **Parallel** relation is formally defined as different steps involving non-overlapping ingredients and utensils. Swapping the ordering of parallel steps will not affect the final dish. The sequential and parallel structure of the steps in a recipe can be characterized as a **flow graph**, where directed edges connect sequential steps (represented as nodes), and the edge direction describes the execution order. All topological sorts of the flow graph will be valid recipes resulting in the same dish. A formal definition will be given in Section 3. The flow graphs of the two recipes *Carnitas Tacos With Cilantro Lime Sauce* and *Carne Asada Tacos* are shown in Figure 1b and 1d respectively.

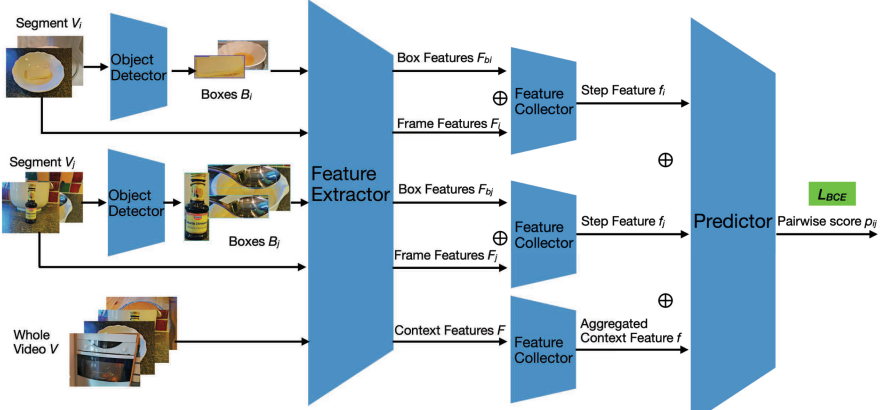
In this paper, we study the problem of predicting the flow graph given a procedural video instance and its step starting and ending timestamps. Some of the challenges associated with predicting flow graphs are: **First**, according to the definition of the parallel relation, the model needs to accurately recognize the involved ingredients and utensils, including distinguishing visually similar utensils and different ingredients. **Second**, cooking involves complex operations that transform ingredients both mechanically and chemically. Attributes such as shape and color can change drastically during this process. Consequently, the model needs to track the state change of the ingredients.

To tackle these challenges, we propose the **Box2Flow** framework, as shown in Figure 2. We first calculate the edge probabilities for all step pairs in a video to get a probability matrix. Then, we create the flow graph from the matrix with a spanning tree algorithm for directed graphs. More specifically, to make our model focus on the ingredients and utensils involved in order to more accurately predict the step relations, we extract the object bounding boxes in the step segments. To tackle the second challenge, we include the whole video as context to monitor the state of each ingredient. We experiment on the labeled MM-ReS[26] and the unlabeled YouCookII[38] datasets. Furthermore, we interpolated the missing frames in MM-ReS to improve the performance. In addition to traditional recall and precision metrics, we use maximal common subgraph[5] for a more structural evaluation. Results show **Box2Flow** can effectively predict the flow graphs.

In summary, the contributions of this paper are:

- We study the less explored problem of generating the flow graph from a single procedural video instance.
- We propose **Box2Flow** to solve the problem of predicting flow graphs from videos. Experiments show the framework effectively predicts the flow graphs.

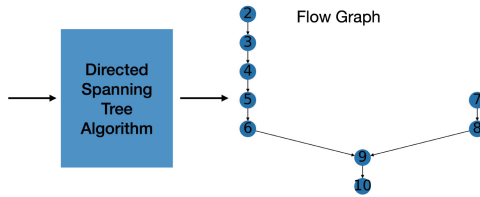
- We interpolate the missing frames in the MM-ReS dataset to extract flow graphs effectively. We also explore the utility of a learned flow graph predictor trained on MM-ReS to a zero-shot transfer task for the unlabeled YouCookII dataset.
- We assess the accuracy of the predicted flow graphs to the ground truth graphs using the structural similarity metric, Maximal Common Subgraphs.



(a) Predicting pairwise relation scores from a pair of video step segments and the whole videos as context.

Probability matrix P

node1\node2	2	3	4	5	6	7	8	9	10
2	-	0.97	0.96	0.96	0.94	0.78	0.81	1	1
3	-	-	0.90	0.98	0.90	0.58	0.86	1	1
4	-	-	-	0.89	0.90	0.10	0.11	1	1
5	-	-	-	-	1	0.73	0.93	1	1
6	-	-	-	-	-	0.15	0.31	0.99	1
7	-	-	-	-	-	-	0.87	1	1
8	-	-	-	-	-	-	-	1	1
9	-	-	-	-	-	-	-	-	1
10	-	-	-	-	-	-	-	-	-



(b) After calculating the edge probabilities for all pairs of steps in a video, we create the flow graph using a spanning tree algorithm for directed graphs.

Fig. 2. Overview of our method. We first predict the edge probabilities for all step segment pairs then create the flow graph using a spanning tree algorithm from the probability matrix.

2 Related Work

Our work relates to video graph representations, flow graph prediction, and their downstream applications. Previous work in computer vision has focused on various graph representations for videos. When characterizing tasks with multiple steps as flow graphs, previous work has focused on flow graph prediction and applications for text, sequences of images, and program codes. However, flow graph prediction from a single procedural video instance has been less explored.

Video graph representations. One of the most studied graph representations is the scene graph. [6,18] generated video scene graphs by predicting the scene graphs at each frame. Although the temporal properties were considered during prediction, the outputs for each frame remain unconnected. [10] added temporal edges between the same node in neighboring frames, which is also called spatio-temporal graph. [31] built a panoptic 3D scene from RGB-D videos by exploiting the actual spatial relations between neighboring scenes. For representing videos as various types of graphs, [14,29] learned a semantic graph from the entire instructional video where the nodes learn semantic concepts and the edges are calculated from the node features. [15,21] learned **general** flow graphs from **multiple** videos for **each task** and [37] built a large flow graph together for all tasks. In these non-instance-based works, if two steps are performed in different orderings in different videos, they are treated as parallel. However, the parallel relationships might not be fully covered due to the limited number of videos in the dataset. On the other hand, the relationships between the steps are intrinsic to the steps themselves through the involved objects. To address the issues caused by task-based methods that learn a single graph from multiple videos, in our work, we focus on an instance-based method that learn flow graphs specific to each available video. This approach allows us to develop more precise and detailed video-specific representations.

Regarding graph downstream applications, [7,24] used scene graphs for video rendering and synthesis, respectively. [25,28] used spatio-temporal graphs for temporal moment localization given language queries and action recognition, respectively. Specifically, [28] included linguistic and visual nodes in their graph. [33] formulated video snippets as graph nodes and snippet correlations as edges for action detection. [2] used task graphs where nodes correspond to objects and edges correspond to actions for video synthesis. [12] represented videos as conjugate task graphs where the nodes are actions, and the edges are states for a single-shot action plan. [13] used graph representations for action segmentation where the nodes are segments, and the edges represent neighboring segment relations. [30] used graphs for video captioning where the nodes include both whole video features and word features.

Flow graph prediction and downstream applications. [22,34] created the fine-grained flow graphs from Japanese and English recipe texts respectively where each node is a named entity. [26,27,36] created recipe step flow graphs from text and images. Specifically, [36] used Japanese language text.

In terms of applications employing flow graphs, [20] used flow graphs from 5G communication base station product manual texts for error detection and correction. [3] used program control flow graphs for malware detection. [8] used **general** task flow graphs for video grounding, which were created from web text. Specifically, only **one** flow graph is created for all videos of the same category. [23] used the help of flow graphs from text to train a captioning model where the inputs include a list of ingredients and a sequence of images where each image is considered as a single step.

In summary, instance-based flow graph prediction from a single video has been less explored. Compared with text, predicting flow graphs from visual inputs are more challenging. The involved objects might not be salient in images and videos. The view points might also change and the cooking process will drastically change the visual appearances of the ingredients, making them difficult to track therefore challenging to create the flow. Meanwhile, long videos contain more information than a sequence of a few frames (less than 100 images) and, subsequently, are more challenging. [26, 27] are image-based methods which average all image features in a single step while we study video-instance inputs. We explicitly model the input images or video clips as sequences to capture the action information. Furthermore, predicting a flow graph for each video can preserve the specific steps in the recipe that might not be covered by the general task flow graph, e.g., A general flow graph of the task `making coffee` might miss some steps specific to certain videos including `add milk foam` and `add syrup` which gives unique flavoring to the recipe. These steps could be included when predicting the flow graph from one video instance. In this paper, we predict a flow graph for each video instance by predicting pairwise edge probabilities from both frame-level and object features, then convert the probability matrix to a flow graph with a spanning tree algorithm.

3 Method

3.1 Flow Graph Definition

A flow graph $F = (S, E)$ is a directed acyclic graph where each node is a step. Let the set of nodes $S = \{S_1, S_2, \dots, S_i, \dots, S_n\}$, where S_i is the i -th step in the recipe. A directed edge (S_i, S_j) exists between if and only if the following rules hold:

1. $i < j$, and
2. S_i and S_j are sequential, and
3. if $j > i + 1$, there is no such k where $i < k < j$, such that both step pairs (S_i, S_k) and (S_k, S_j) are sequential.

Rule 1 determines the graph’s flow where the later nodes are descendants. Rule 2 considers the sequential relation as edges, and rule 3 does not allow skip edges.

In other words, an edge connects a step with its direct consequence. If S_i and S_j are sequential but indirect (e.g., steps 2 and 9 in `Carnitas Tacos With Cilantro Lime Sauce Recipe`, as shown in Figure 1a, 1b), there exists a path with length at least two between S_i and S_j and vice versa.

3.2 Pairwise Edge Probability

Given a video V and the start and end frames for each step $T = \{(s_i, e_i) | 1 \leq i \leq n\}$, our goal is to predict the flow graph F or the set of edges $E \subseteq \{(S_i, S_j) | 1 \leq i < j \leq n\}$. In addition, we denote the i -th video segment $V[s_i : e_i]$ as V_i .

To predict the pairwise edge probability, we first extract the object bounding boxes with an object detector:

$$B_i = F_{od}(V_i) \quad (1)$$

where $B_i = \{(x_{min_{kt}}, y_{min_{kt}}, x_{max_{kt}}, y_{max_{kt}})\} \cdot (x_{min_{kt}}, y_{min_{kt}})$ is the top-left corner and $(x_{max_{kt}}, y_{max_{kt}})$ is the bottom-right corner of the k -th box in the t -th frame of the segment. The frame patches defined by B_i are denoted as $V_i[B_i]$.

Next, we extract the object features with a video encoder:

$$F_{b_i} = F_{fe}(V_i[B_i]) \quad (2)$$

where $F_{b_i} \in \mathbb{R}^{K_i \times d}$. $K_i = \sum_{t=1}^{e_i - s_i + 1} k_t$ is the total number of bounding boxes in the segment and d is the output feature dimension.

Similarly, the frame features of V_i can be extracted as

$$F_i = F_{fe}(V_i) \quad (3)$$

where the bounding boxes can be treated as $(1, 1, W, H)$ for all frames. W is the frame width and H is the frame height. $F_i \in \mathbb{R}^{(e_i - s_i + 1) \times d}$.

As the step relations given only two video segments can be ambiguous, we also include the whole video feature F as context, which consists of all frame-level features stacked together:

$$F = \text{stack}(F_1, F_2, \dots, F_i, \dots, F_n) \quad (4)$$

$F \in \mathbb{R}^{N \times d}$ where $N = \sum_{i=1}^n (e_i - s_i + 1)$ is the total number of non-background frames related to the task.

Then, we aggregate the frame and box features for each segment using BERT with adapters[11]. The features are first projected to BERT input embedding dimension through the same linear layer:

$$F_{e_i} = \tanh(F_{fc}(F_i)) \quad (5)$$

$$F_{e_{b_i}} = \tanh(F_{fc}(F_{b_i})) \quad (6)$$

$$F_e = \tanh(F_{fc}(F)) \quad (7)$$

The frame and the box embeddings are stacked together with the BERT [CLS] embedding and fed through the transformer encoder to extract the step features. The position IDs for the k_t boxes and the frame embeddings in the t -th frame are all t , and the [CLS] embedding has position ID 0. The output of [CLS] representation is taken as the aggregated feature. For the context feature, only the frame feature F is used.

$$f_i = F_{bert_{step}} \left(\text{stack}(F_{cls}, F_{e_i}, F_{e_{b_i}}) \right) \quad (8)$$

$$f = F_{bert_{ctx}} \left(\text{stack}(F_{cls}, F_e) \right) \quad (9)$$

where $F_{bert_{step}}, F_{bert_{ctx}}$ are two different BERT adapters for step features and context features respectively. F_{cls} is BERT [CLS] embedding. $f_i, f \in \mathbb{R}^{d_{bert}}$ are 1-D vectors with BERT output dimension.

Finally, two-step features $f_i, f_j, (i < j)$ and the context feature f are concatenated and fed through an MLP to predict the pairwise sequential probability:

$$p_{ij} = \sigma(F_{mlp}(f_i \oplus f_j \oplus f)) \quad (10)$$

where $\sigma(\cdot)$ is the Sigmoid function and \oplus stands for concatenation.

During training, we use the weighted binary cross-entropy loss:

$$L = -w_s \sum_{1 \leq i < j \leq n} [w_p y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij})], \quad (11)$$

where w_s is the video sample weight, w_p is the positive weight for unbalanced label distribution. $y_{ij} = 1$ for sequential relation, both direct and indirect, and $y_{ij} = 0$ for parallel relation.

Multi-modality. Our framework can be easily extended to prediction with only text or both video and text modalities. When only the recipe text is available, we have $R = R_1 \oplus R_2 \oplus \dots \oplus R_i \oplus \dots \oplus R_n$, where R_i is the text for the i -th step. Then, the text tokens are directly fed to the BERT adapters to get the features for the MLP module in Equation 10:

$$f_{text_i} = F_{bert_{text}}(R_i) \quad (12)$$

$$f_{text} = F_{bert_{text}}(R) \quad (13)$$

The text-described step and context features share the same adapter. The sequential probability is calculated as follows:

$$p_{ij} = \sigma(F_{mlp}(f_{text_i} \oplus f_{text_j} \oplus f_{text})) \quad (14)$$

When both video and text are available, after the features $f_i, f, f_{text_i}, f_{text}$ are extracted as Equation 8, 9, 12, 13, the probability is calculated as:

$$p_{ij} = \sigma(F_{mlp}(f_i \oplus f_j \oplus f \oplus f_{text_i} \oplus f_{text_j} \oplus f_{text})) \quad (15)$$

Three different adapters are involved: video step, video context and text.

3.3 Graph Construction

We construct the flow graph after all the pairwise scores $P = \{p_{ij} | 1 \leq i < j \leq n\}$ have been calculated. Since most of the flow graphs in the real world are trees with at most one descendent for each node, we focus on constructing trees. Because of rule 3 in Section 3.1, if the flow graph is a tree and there is an edge between (S_i, S_j) , S_j has to be the earliest step such that S_i and S_j are sequential.

Therefore, we first select the edges according to the standard probability threshold 0.5 to get a set of candidate edges $E_{can} = \{(S_i, S_j) | p_{ij} > 0.5\}$. Then, for each step i , we select the earliest step j such that $(S_i, S_j) \in E_{can}$ to form the flow graph, as in Algorithm 1.

4 Experiments

4.1 Datasets and Metrics

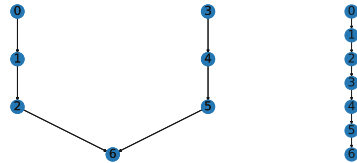
We use two datasets, labeled MM-ReS[26] and unlabeled YouCookII[38].

MM-ReS consists of recipe texts, step images and annotated flow graphs. The original dataset includes 9850 recipes collected from the Internet. Since we focus on predicting tree graphs, we only use the 8370 recipes with tree annotations and randomly split into training, validation and test sets, with 6696, 837, 837 each. 64.5k steps are annotated in flow graphs. The dataset includes 131k images, with 2.8 images/step on average for steps with images. The images can be treated as short video clips for each step. Meanwhile, 18% of the steps do not have images and need to be removed, zero-padded or interpolated.

YouCookII consists of cooking videos from YouTube, with step starting and ending time annotations and step text descriptions rephrased by annotators but without annotated flow graphs. We use 1181 training videos and 414 validation videos, which are still available on Youtube. There are 7.7 steps/video on average. For evaluation, we manually annotated the flow graphs of 39 videos in the training set and 63 videos in the validation set. 97 of the annotations are trees. During training, we combined the manual annotation and predictions from a text model trained on MM-ReS as labels.



(a) Peanut Butter and Jelly Sandwich.



(b) Ground truth flow graph.

(c) Predicted flow graph.

Fig. 3. An example of ground truth and predicted flow graphs where the recall and precision are high but very different structurally. The recipe is **Peanut Butter and Jelly Sandwich** from MM-ReS.

Algorithm 1. Flow graph from probability matrix

Inputs: probability matrix P , number of steps n

Output: edge set E

$E \leftarrow \phi$

$E_{can} \leftarrow \{(S_i, S_j) | p_{ij} > 0.5\}$

for $i = 1$ to n **do**

$I \leftarrow \{j | (S_i, S_j) \in E_{can}\}$

if $I \neq \phi$ **then**

$j \leftarrow \operatorname{argmin}_{j>i} (j \in I)$

$E \leftarrow E \cup \{(S_i, S_j)\}$

end if

end for

Return E

Metrics. Following [26], we report edge-level recall R_e , precision P_e , F1 and recipe-level recall R_r , precision P_r , F1, which are calculated as the follows:

Suppose the ground truth edge set of the i -th video in the dataset is E_i , the predicted edge set is \hat{E}_i , $|\cdot|$ denotes set cardinality and there are M videos in the dataset,

$$R_e = \frac{\sum_{i=1}^M |E_i \cap \hat{E}_i|}{\sum_{i=1}^M |\hat{E}_i|}, \quad P_e = \frac{\sum_{i=1}^M |E_i \cap \hat{E}_i|}{\sum_{i=1}^M |E_i|} \quad (16)$$

The edge-level F_1 is the harmonic average between R_e and P_e . Define R_i, P_i as the precision and recall for each recipe, calculated as

$$R_i = \frac{|E_i \cap \hat{E}_i|}{|\hat{E}_i|}, \quad P_i = \frac{|E_i \cap \hat{E}_i|}{|E_i|} \quad (17)$$

Define $F1_i$ as the harmonic average between R_i and P_i . Then $R_r, P_r, F1_r$ are calculated as $\frac{1}{M} \sum_{i=1}^M R_i, \frac{1}{M} \sum_{i=1}^M P_i, \frac{1}{M} \sum_{i=1}^M F1_i$ respectively.

However, these metrics might not accurately reflect the structural similarity between the predicted and the ground truth graph, as shown in Figure 3, taken from **Peanut-Butter-and-Jelly-Sandwich-1** recipe in MM-ReS. The ground truth shows two branches merging, while the predicted is a chain. Therefore, the structures are very different. However, the only different edge is (2,6) in the ground truth and (2,3) in the predicted. Recall, precision, and F1 are all as high as 83% in this case. As a result, we also include a structural similarity metric maximal common subgraph (MCS)[5]. Define cc as the number of nodes in the connected component of $E \cap \hat{E}$ with the maximum size and n is the number of nodes in E , $MCS = cc/n$. In the example, the maximal common subgraph is $\{(3,4),(4,5),(5,6)\}$ with 4 nodes and $mcs=4/7=57\%$.

4.2 Compared Methods

We investigate methods using different modalities, including video-only, text-only, and video+text. ¹Specifically since only a few annotations are available for YouCookII, we directly transfer a pre-trained model on MM-ReS for text-only to show zero-shot ability.

We include video captioning as a baseline for video-only methods. Captions are first generated for videos then flow graphs are created from the captions using Equation 14. We also manually annotated some examples from generated captions. The details are in our supplement. We compare with the baseline video captioning methods **MART**[17] and **VLTinT**[35].

To show the effects of bounding boxes, we compare **Box2Flow** with its variance using only frame features F_{e_i} in Equation 8 but not box features $F_{e_{b_i}}$. The variance is denoted by "f". To remove the effects of more parameters introduced

¹ Our results are not directly comparable with [26, 27] because of different evaluation subsets and code not available.

by two adapters, we also include models using the same adapter for both context and step features, denoted by "1". Specifically, for video+text methods on YouCookII, all methods are trained with one adapter.

For MM-ReS, we fix bottom-up attention[1] for feature extraction and compare two different object detectors, Detectron2[32] pre-trained on COCO[19] and SAM[16] masks, denoted by "C" and "S" respectively. As 18% of the steps do not have images in MM-ReS, we also study the effect of interpolating the missing images using instruct-pix2pix[4] for image+text models, denoted by "i". Otherwise, we zero-pad the image features for image+text methods and directly remove these nodes for image-only methods.

For YouCookII, we compare two different frame feature extractors, Densecap [39] and SlowFast[9], denoted by "D" and "SF" respectively. Specifically, only SlowFast can include bounding boxes for feature extraction. We fix Detectron2 as the object detector.

We also include a naive **chain** baseline which only requires the number of steps: $E = \{(S_1, S_2), (S_2, S_3), \dots, (S_i, S_i + 1), \dots, (S_{n-1}, S_n)\}$.

The implementation details are in our Supplement.

4.3 Results and Evaluation

Table 1 shows the results on MM-ReS dataset. For image-only methods, we remove the steps without images during training and evaluation. The flow graphs will change after step removal. When removing steps from the graph, a node is directly deleted if it has no ancestor or descendant. Otherwise, its ancestor is directly connected to its descendant. To enable comparison across modalities, we also evaluate text-only and chain methods on steps with images, denoted by *. We remove nodes without images from text-only model and chain outputs using the above process. Table 2 shows the results on YouCookII dataset.

Effects of Modalities. For video or image only methods, only Box2Flow-SF surpassed the naive chain baseline in Table 1 and 2, showing that directly predicting flow graphs from videos is a challenging problem. We show in our Supplement that the flow graphs predicted by the text model from generated captions do not accurately capture the true structure of the captions. The video captions are inconsistent in ingredients and the scores of manually labeled caption flow graphs would be much lower. Therefore, directly predicting flow graphs from videos is needed. Using text modalities can significantly improve the performance, surpassing the chain baselines. The text model trained on MM-ReS also shows zero-shot ability, achieving high performance on the YouCookII dataset. This is not to be taken for granted, as video clip descriptions are different from formal recipe steps. Furthermore, the step texts in MM-ReS are significantly longer than those in YouCookII. Each step has an average of 32.3 BERT tokens in MM-ReS, while YouCookII only has 14.4 tokens. Yet videos provide complementary information. Methods using both video and text improve upon text-only models. For example, in Figure 1c, 1d, the text mentions that step 3, "heat up the tortillas", is parallel with steps 2 and 4 as step 3 introduces a new ingredient.

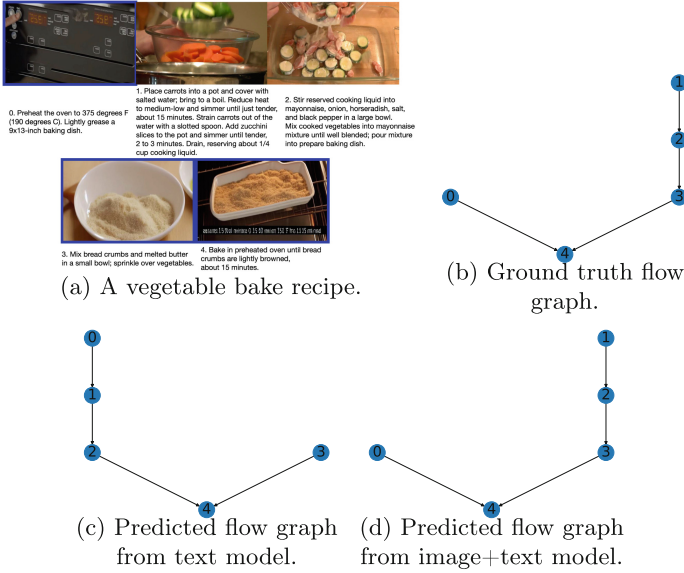


Fig. 4. An example from MM-ReS. The text-only model did not predict the graph correctly, while the image+text model did. The interpolated images are marked in blue.

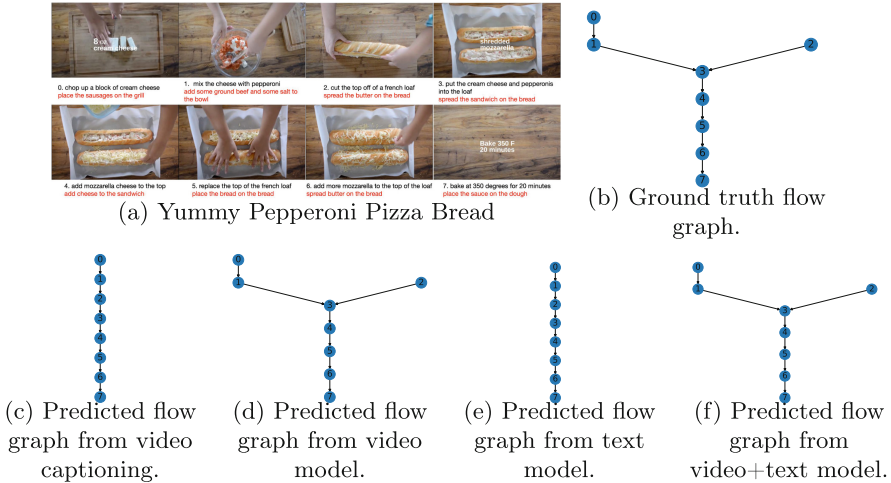


Fig. 5. An example from YouCookII. The predicted captions are in red in (a). Video captioning and the text model did not predict the graph correctly, while the video and video+text models did.

However, the video shows the meat is grilled first; then, the tortillas are added while the meat is still grilling and share the same grill, showing the steps are actually sequential.

Effects of Bounding Boxes. Comparing image-only methods and image+text methods in Table 1, video-only methods and video+text methods in Table 2, where step and context features share the same adapter for methods using bounding boxes, the results show using bounding boxes can improve the performance by focusing on the involved ingredients and utensils.

Effects of Object Detectors, Feature extractors and Interpolation. Table 1 shows SAM masks can further improve the performance from Detectron2 object detectors on COCO. Instruct-pix2pix interpolation improves on the COCO detector more than SAM. Table 2 shows video-only frame-level SlowFast is better than Denscap features in structures. Including bounding boxes further improves the performance when using SlowFast feature extractor.

In summary, `Box2Flow` can predict the flow graphs effectively and can be used together with different object detectors and feature extractors. Videos can provide information that complements text, and bounding boxes can further improve the effectiveness, even without introducing more parameters in the model. We include more ablation studies, including the effects of context features, SAM mask selection and binary vs soft labels in our Supplement.

4.4 Qualitative Results

We show some recipe examples and the flow graph predictions of various methods. The examples are selected based on the largest MCS improvement using video + text from text-only. We include more examples in our Supplement.

Figure 4 shows an example from MM-ReS where the text model did not predict the correct graph, but the image+text model did. The shared edges between the ground truth and the text graph are $\{(1,2),(3,4)\}$; therefore, the precision, recall, and F1 of the recipe are all $2/4=0.5$. The maximum common subgraph is 1-2 or 3-4; therefore, $MCS=2/5=0.4$. For the image+text model, all metrics will be 1. The text model treats step 0 preheat oven and step 1 cook vegetables in a pot as sequential; step 2 mix vegetables and step 3 sprinkle bread crumbs on vegetables as parallel, showing the model did not notice some word details from the long instructions. The images from step 0, 3, 4 are originally missing from the dataset and are interpolated with instruct-pix2pix, marked by blue borders. The interpolated images correctly show the oven and baking action in steps 0 and 4 and the bread crumbs in step 3, although missing the vegetables. The image+text model still correctly determined the edge (0,4) and the step 2, 3 should be sequential.

Table 1. MM-ReS results in percentage. The best performance evaluated on all nodes is marked **bold**. * means evaluation on steps with images only.

Modality	Method	Edge Recall	Edge Precision	Edge F1	Recipe Recall	Recipe Precision	Recipe F1	MCS
Images	MART[17]	80.0	81.0	80.5	81.5	82.5	81.8	77.6
	VLTinT[35]	81.8	82.6	82.2	82.2	82.9	82.4	78.6
	Box2Flow-f	80.5	81.0	80.7	81.5	82.3	81.7	77.7
	Box2Flow-1	80.1	80.6	80.3	81.9	82.3	82.0	78.4
Text	Box2Flow	82.5	83.0	82.8	87.3	87.5	87.4	81.7
	Box2Flow*	82.9	83.3	83.1	86.1	86.3	86.1	81.7
Images+Text	Box2Flow-f	83.6	83.9	83.7	87.2	87.5	87.3	82.1
	Box2Flow-C	83.7	83.9	83.8	87.3	87.5	87.4	82.8
	Box2Flow-Ci	84.5	84.7	84.6	87.7	87.9	87.8	83.1
	Box2Flow-S	84.7	84.9	84.8	87.9	88.1	88.0	83.3
	Box2Flow-S1	83.4	83.6	83.5	87.1	87.3	87.2	82.7
	Box2Flow-Si	84.6	84.8	84.7	87.9	88.0	87.9	83.4
-	chain	84.1	84.1	84.1	83.4	83.4	83.4	78.6
	chain*	85.1	85.0	85.0	84.5	84.3	84.3	80.5

Table 2. YouCookII results in percentage. The best performance is marked **bold**.

Modality	Method	Edge Recall	Edge Precision	Edge F1	Recipe Recall	Recipe Precision	Recipe F1	MCS
Video	MART[17]	74.3	76.0	75.2	77.6	78.9	78.1	70.5
	VLTinT[35]	73.3	75.6	74.4	75.2	76.4	75.7	67.2
	Box2Flow-fD	78.6	79.1	78.8	80.0	80.0	80.0	69.6
	Box2Flow-ISF	77.0	77.8	77.4	78.9	80.1	79.3	71.8
	Box2Flow-SF1	79.3	79.3	79.3	80.7	80.9	80.7	72.0
	Box2Flow-SF	80.5	80.3	80.4	81.9	81.5	81.7	72.9
Text	Box2Flow-MMRs	85.5	85.8	85.6	87.8	88.0	87.9	80.8
Video+Text	Box2Flow-fD	85.8	86.0	85.9	87.9	87.9	87.9	81.7
	Box2Flow-ISF	85.5	85.5	85.5	88.1	88.0	88.0	81.8
	Box2Flow-SF	86.3	86.5	86.4	88.5	88.7	88.6	81.8
-	chain	81.0	80.8	80.9	82.9	82.5	82.7	72.7

Figure 5 shows Yummy Pepperoni Pizza Bread recipe from YouCookII. Video captioning from MART and text model did not predict the correct graph, but video and video+text model did. Both video captioning and the text model predicted chains. The only different edge between the ground truth and the chain is (1,3) in ground truth and (1,2) in the chain; therefore, the precision, recall, and F1 of the recipe are all $6/7=0.875$. The maximum common subgraph is the part from 2-7; therefore, $MCS=6/8=0.75$. For the video and video+text model, all metrics will be 1. MART did not generate the correct ingredients for the first two steps and recognized cheese as butter in step 6. It also generated the impossible action "spread sandwich on bread" in step 3. The captions for steps 4 and 5 are correct. The text model predicted the generated captions as a chain even with inconsistent ingredients throughout the recipe. Meanwhile, the video model and the video+text model correctly predicted step 2 is parallel to the previous steps from the visual clue, correcting the mistake made by the text model.

5 Conclusion

We have studied the less explored problem, predicting the flow graph from a single procedural video instance. We proposed `Box2flow` framework, which exploits the bounding boxes and creates a spanning tree from pairwise sequential probabilities. Although a challenging problem, `Box2flow` can predict the flow graphs effectively. This also opens up possible future research directions: predicting the flow graphs more effectively from video or image-only features and exploring their utility in downstream applications, like more structured video captioning and planning.

References



1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
2. Bar, A., Herzig, R., Wang, X., Rohrbach, A., Chechik, G., Darrell, T., Globerson, A.: Compositional video synthesis with action graphs. arXiv preprint [arXiv:2006.15327](https://arxiv.org/abs/2006.15327) (2020)
3. Bobrovnikova, K., Lysenko, S., Savenko, B., Gaj, P., Savenko, O.: Technique for iot malware detection based on control flow graph analysis. *Radioelectronic and Computer Systems* **1**, 141–153 (2022)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Bunke, H., Shearer, K.: A graph distance metric based on the maximal common subgraph. *Pattern Recogn. Lett.* **19**(3–4), 255–259 (1998)
6. Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., Yang, M.Y.: Spatial-temporal transformer for dynamic scene graph generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16372–16382 (2021)
7. Cong, Y., Yi, J., Rosenhahn, B., Yang, M.Y.: Ssgvs: Semantic scene graph-to-video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2554–2564 (2023)
8. Dvornik, N., Hadji, I., Pham, H., Bhatt, D., Martinez, B., Fazly, A., Jepson, A.D.: Graph2vid: Flow graph to video grounding for weakly-supervised multi-step localization. arXiv preprint [arXiv:2210.04996](https://arxiv.org/abs/2210.04996) (2022)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
10. Holm, F., Ghazaei, G., Czempiel, T., Özsoy, E., Saur, S., Navab, N.: Dynamic scene graph representation for surgical video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 81–87 (2023)
11. Houlisby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
12. Huang, D.A., Nair, S., Xu, D., Zhu, Y., Garg, A., Fei-Fei, L., Savarese, S., Niebles, J.C.: Neural task graphs: Generalizing to unseen tasks from a single video demonstration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8565–8574 (2019)

13. Huang, Y., Sugano, Y., Sato, Y.: Improving action segmentation via graph-based temporal reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14024–14034 (2020)
14. Hussein, N., Gavves, E., Smeulders, A.W.: Videograph: Recognizing minutes-long human activities in videos. arXiv preprint [arXiv:1905.05143](https://arxiv.org/abs/1905.05143) (2019)
15. Jang, Y., Sohn, S., Logeswaran, L., Luo, T., Lee, M., Lee, H.: Multimodal subtask graph generation from instructional videos. arXiv preprint [arXiv:2302.08672](https://arxiv.org/abs/2302.08672) (2023)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
17. Lei, J., Wang, L., Shen, Y., Yu, D., Berg, T.L., Bansal, M.: Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. arXiv preprint [arXiv:2005.05402](https://arxiv.org/abs/2005.05402) (2020)
18. Li, Y., Yang, X., Xu, C.: Dynamic scene graph generation via anticipatory pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13874–13883 (2022)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
20. Luo, R., Zhu, Q., Chen, Q., Wang, S., Wei, Z., Sun, W., Tang, S.: Operation diagnosis on procedure graph: The task and dataset. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3288–3292 (2021)
21. Mao, W., Desai, R., Iuzzolino, M.L., Kamra, N.: Action dynamics task graphs for learning plannable representations of procedural tasks. arXiv preprint [arXiv:2302.05330](https://arxiv.org/abs/2302.05330) (2023)
22. Mori, S., Maeta, H., Yamakata, Y., Sasada, T.: Flow graph corpus from recipe texts. In: LREC. pp. 2370–2377 (2014)
23. Nishimura, T., Hashimoto, A., Ushiku, Y., Kameko, H., Yamakata, Y., Mori, S.: Structure-aware procedural text generation from an image sequence. *IEEE Access* **9**, 2125–2141 (2020)
24. Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F.: Neural scene graphs for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2856–2865 (2021)
25. Ou, Y., Mi, L., Chen, Z.: Object-relation reasoning graph for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20133–20142 (2022)
26. Pan, L.M., Chen, J., Wu, J., Liu, S., Ngo, C.W., Kan, M.Y., Jiang, Y., Chua, T.S.: Multi-modal cooking workflow construction for food recipes. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1132–1141 (2020)
27. Pan, L., Chen, J., Liu, S., Ngo, C.W., Kan, M.Y., Chua, T.S.: A hybrid approach for detecting prerequisite relations in multi-modal food recipes. *IEEE Trans. Multimedia* **23**, 4491–4501 (2020)
28. Rodriguez-Opazo, C., Marrese-Taylor, E., Fernando, B., Li, H., Gould, S.: Dori: Discovering object relationships for moment localization of a natural language query in a video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1079–1088 (2021)

29. Schiappa, M.C., Rawat, Y.S.: Svgraph: Learning semantic graphs from instructional videos. In: 2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM). pp. 45–52. IEEE (2022)
30. Tu, Y., Zhou, C., Guo, J., Li, H., Gao, S., Yu, Z.: Relation-aware attention for video captioning via graph learning. *Pattern Recogn.* **136**, 109204 (2023)
31. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7515–7525 (2021)
32. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
33. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10156–10165 (2020)
34. Yamakata, Y., Mori, S., Carroll, J.A.: English recipe flow graph corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 5187–5194 (2020)
35. Yamazaki, K., Vo, K., Truong, Q.S., Raj, B., Le, N.: Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In: Proceedings of the AAAI Conference on Artificial intelligence. vol. 37, pp. 3081–3090 (2023)
36. Zhang, Y., Yamakata, Y., Tajima, K.: Miais: a multimedia recipe dataset with ingredient annotation at each instructional step. In: Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and related APplications. pp. 49–52 (2022)
37. Zhou, H., Martín-Martín, R., Kapadia, M., Savarese, S., Niebles, J.C.: Procedure-aware pretraining for instructional video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10727–10738 (2023)
38. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI Conference on Artificial Intelligence. pp. 7590–7598 (2018), <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344>
39. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8739–8748 (2018)
40. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3537–3545 (2019)



Learning Geometry of Pose Image Manifolds in Latent Spaces Using Geometry-Preserving GANs

Shenyuan Liang¹, Benjamin Beaudett¹ , Pavan Turaga², Saket Anand³,
and Anuj Srivastava¹ 

¹ Florida State University, Tallahassee, FL 32309, USA
{s120fu, bbeaudett, asrivastava}@fsu.edu

² Arizona State University, Tempe, AZ 85281, USA
pturaga@asu.edu

³ IIT-Delhi, Delhi, India
anands@iiitd.ac.in

Abstract. The goal of this paper is to learn the differential geometry of pose image manifolds for 3D objects. Indexed by the rotation group $SO(3)$, a pose manifold constitutes images of a 3D object from all viewing angles. Learning geometry implies computing geodesics, intrinsic statistics (means, etc), and curvatures on estimated manifolds. As these goals are unattainable in the huge image space, we perform dimension reduction that is **geometry preserving** and **invertible**. This paper introduces two distinct concepts: (1) A **Geometry-Preserving StyleGAN** (GP-StyleGAN2) that maps training images to a low-dimensional latent space with two novel geometry-preserving terms. These terms penalize changes in pairwise distances between points and pairwise angles between tangent spaces under the map. (2) Densifying the estimated manifold in latent space using **Euler's Elasticae**-based nonlinear interpolations between sparse data points. In contrast to the past findings, the latent pose manifolds are found to be distinctly nonlinear and similar in shape across objects. Incorporating these features results in superior performance in image interpolation, denoising, and computing image summaries when compared to state-of-the-art GANs and VAEs.

Keywords: Manifold Learning · Pose Image Manifold · Elasticae · Latent Space Geometry · Geodesics · Geometric GAN

1 Introduction

Image manifolds are subsets of image spaces corresponding to images of 3D objects of interest. In this paper, we focus on specific image manifolds called *rotation or pose manifolds*. A pose manifold is the set of images of an object under all 3D rotations (while fixing other imaging conditions). Even though images are high-dimensional, the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78398-2_4.

pose manifolds are typically low-dimensional and are nonlinearly embedded in the huge ambient Euclidean space. Learning the differential geometries of image manifolds has been a long-standing and challenging goal in the field, especially when using limited training data. Learning, characterizing, and exploiting this geometry can help accomplish several goals in image analysis and computer vision: interpolating between images using geodesics, denoising an image using manifold projection, making statistical analysis more interpretable by adhering to the manifold structure, and creating simple yet powerful generative and discriminative models by defining probability distributions on the manifold. The goal of this paper is to *learn the differential geometry of image manifolds for individual 3D objects*, enabling computation of geodesics, tangents, curvatures, and intrinsic statistics (means, etc.) on the estimated manifolds.

We will follow the notation in [13] to develop a mathematical formulation. Let α be a 3D object, such as a chair, airplane, or car, and let O^α denote its 3D geometry and reflectance model. Let $s \in SO(3)$ represent the 3D pose of O^α relative to the camera and \mathbb{P} be the orthographic projection of sO^α into the focal plane of the camera, resulting in an $n \times n$ image $\mathbb{P}(sO^\alpha)$. As mentioned earlier, all other imaging variables, e.g. illumination, are fixed for this discussion.

Definition 1 *Under these conditions, the set $\mathcal{I}^\alpha = \{\mathbb{P}(sO^\alpha) \in \mathbb{R}^{n \times n} | s \in SO(3)\}$, a subset of $\mathbb{R}^{n \times n}$, is called the rotation or pose manifold of α .*

Using additional assumptions on the smoothness of \mathbb{P} and non-symmetry of O^α with respect to the rotation group, the set \mathcal{I}^α shares the closed, boundary-free manifold topology of $SO(3)$. (This statement deserves additional consideration to be precise but we leave the details for a future paper.) It is thus three-dimensional and its nonlinear embedding is small (almost singular) in the ambient space $\mathbb{R}^{n \times n}$.

Problem Specification: Given a training set of rotation-image pairs $\mathcal{R}_{train} = \{(s_i, I_i = \mathbb{P}(s_i O^\alpha)) \in SO(3) \times \mathcal{I}^\alpha\}_{i=1}^m$, our goal is to estimate the manifold \mathcal{I}^α . We want not only to learn the topological set \mathcal{I}^α , but also to characterize its local and global geometry. This characterization will enable us to compute quantities on the manifold such as geodesics, geodesic distances, and intrinsic statistical summaries. The task of learning \mathcal{I}^α from \mathcal{R}_{train} is challenging because (i) Learning nonlinear manifolds typically requires large sample sizes that are even greater in high-dimensional (n^2) spaces, and (ii) The underlying geometry of \mathcal{I}^α is often complex and does not follow known simplifications such as spheres, ellipsoids, or hyperbolic spaces.

The difficulties associated with high dimensions can be mitigated by mapping to a smaller Euclidean space $\mathbb{R}^d, d \ll n^2$. If we have a map $\Phi : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^d$ which is approximately isometric and invertible on the manifold, we can analyze the geometry on the simpler latent manifold $\mathcal{M}^\alpha = \Phi(\mathcal{I}^\alpha) \subset \mathbb{R}^d$ using $\{(s_i, \Phi(I_i))\}_{i=1}^m$ and map the results back to the image manifold using Φ^{-1} . The next question is: What is a good choice of Φ ? Existing methods for manifold learning (such as LLE, Isomap, t-SNE, etc.), dimension reduction and latent space representation either distort the geometry of \mathcal{I}^α or are not invertible (see section 2). We require a new technique.

Our Approach: There are two main elements to our approach. Firstly, we seek a mapping Φ that preserves the geometry of the pose manifold. Secondly, rather than assuming

a flat geometry in the latent space as is often the case in the current literature, we use nonlinear interpolations between mapped points in \mathbb{R}^d to discover the unknown manifold. We introduce these items here and elaborate on the details in section 3.

(1) Geometry-Preserving and Invertible Dimension Reduction: How can one design a Φ that maximally preserves the geometry? Preserving the geometry implies that distances, angles, and curvatures remain similar from the domain to the range of Φ . We create Φ by training with a neural network that combines elements of GANs and autoencoders, specifically a modified version of AE-StyleGAN2. The decoder trained as part of this model provides an inverse map as well. We build on this architecture by introducing two geometry-preserving loss terms, and call our model *Geometry-Preserving StyleGAN2* or GP-StyleGAN2. The new terms help preserve (i) pairwise Euclidean distances between point locations, and (ii) pairwise dissimilarity (a measure of local curvature) between tangent space orientations. Here we use all training points to compute pairwise distances, and not just the neighbors, and treat Euclidean distances as *extrinsic* distances to help learn the global geometry. Estimation of tangent planes $\{T_i\}$ at points $\{I_i\}$ used to compute orientation dissimilarity is described later.

(2) Discovering the Manifold using Nonlinear Elasticae: We use the trained Φ to map points from another sparse set $\mathcal{R}_{test} \subset \mathbb{R}^{n^2}$ (disjoint from \mathcal{R}_{train}) to the latent space, resulting in $\{\Phi(I_i) \in \mathbb{R}^d\}$. These points lie on the latent manifold \mathcal{M}^α and we can use them to uncover it in more detail. We do this by interpolating between neighboring mapped points using *free elasticae* [25, 29]. Elasticae use curved interpolations between pairs $(\Phi(I_i), \Phi(I_j))$, with curvatures dictated by their distance and the misalignment of tangent planes $(d\Phi(T_i), d\Phi(T_j))$. Repeatedly applying this tool between neighboring points, we ‘fill in’ the manifold with arbitrarily dense point sets and produce an estimated latent manifold $\widehat{\mathcal{M}}^\alpha$ and the corresponding image manifold $\widehat{\mathcal{I}}^\alpha = \Phi^{-1}(\widehat{\mathcal{M}}^\alpha)$.

Knowing the geometry of \mathcal{M}^α allows us to improve performance in some crucial vision tasks including: (1) Image Interpolation Using Geodesics: Given any two images $I_i \equiv \mathbb{P}(s_i O^\alpha)$ and $I_j \equiv \mathbb{P}(s_j O^\alpha)$ in the test data, the task is to estimate the image path $t \mapsto \mathbb{P}(x(t) O^\alpha)$, where $x : [0, 1] \rightarrow SO(3)$ is a geodesic between s_i, s_j in $SO(3)$. The image path should resemble the video of a rotating object. (2) Intrinsic Image Statistics and Modeling: Given a set of images $\{I_i\}$, one would like to compute their summary statistics (mean, covariance) and develop statistical models as elements of \mathcal{I}^α rather than \mathbb{R}^{n^2} . (3) Image Denoising Using Manifold Projection: Given a noisy or corrupted image $J \in \mathbb{R}^{n^2}$ known to be associated with an $I \in \mathcal{I}^\alpha$, we seek a tool to denoise it.

2 Related Works

In recent years, deep neural networks (DNNs) have provided powerful tools for encoding of images by mapping them to low-dimensional latent spaces. In the following we summarize some past ideas that are most relevant to our method.

Manifold Learning Techniques: From the pre-deep network era, there is a long list of learning methods that sought nonlinear dimension-reduction while preserving some

geometric properties for image data, including LLE [33], Isomap [40], LTSA [44], Laplacian eigenmaps [3], Hessian eigenmaps [10], diffusion maps [7], vector diffusion maps [37], Riemannian relaxation [27], and t-SNE [26]. These were successful in mapping image data into smaller Euclidean spaces while preserving pairwise distances or other local properties. However, these mostly only go as far as embedding the observed training points in a low-dimensional space. They fall far short of our goal of creating an invertible map that can read out-of-sample points in both the input and latent spaces. Several recent papers such as FM-VAE [6], IRVAE [42], GRAE [11], structure-preserving AE [38], GGAE [24], and DIMAL [31] seek geometry-aware manifold learning using different DNN architectures. Our work constitutes further exploration of this area.

Differential Geometry of Latent Spaces: Bengio et al. [4] stressed the importance of understanding the geometry of latent space representations. Several papers [23, 35, 36] have investigated this geometry, mainly utilizing existing architectures geared towards image synthesis, and reported them to be (surprisingly) flat. Other papers [2, 9, 16] demonstrated that Jacobian-based Riemannian metrics on the latent space produce better inference results than using Euclidean distance. Sáez et al. [30] fitted local, constant-curvature patches to data using Gromov-Hausdorff distance and Bayesian optimization. Zhang and Jiang [43] presented a method for geometric space selection in representation learning, allowing data points to select optimal geometric spaces.

GANs and VAEs: Goodfellow et al. [12] introduced the basic framework and training procedure for generative adversarial networks (GANs). Radford et al. [32] improved their stability and efficiency, while Karras et al. [18] proposed Progressive Growing GAN and Style-Based GAN that incorporated regularizations. Han et al. [14] designed AE-StyleGAN2 for more disentangled latent space and improved efficiency. Kingma and Welling [22] introduced variational autoencoder (VAE) to map input data to a low-dimensional latent space (encoder) and back (decoder). VAEs have been extended in various directions. Davidson et al. [8] proposed the *Hyperspherical VAE* (SVAE) that samples latent vectors on a unit sphere. Chadebec and Allasonière [5] proposed the geometry-based *Riemannian Hamiltonian VAE* (RHVAE), which models the latent space as a Riemannian manifold, combining Riemannian metric learning and geodesic shooting. Huh et al. [17] proposed Quotient VAEs.

3 Proposed Framework

In this section, we present the design of *Geometry Preserving StyleGAN2* (GP-StyleGAN2), which facilitates learning by preserving the geometry of the image manifold. We start with a brief introduction to StyleGANs and AE-StyleGAN2. Consider a set of images $\{I_i \in \mathcal{I}\}_{i=1}^b$ and random latent vectors $\{v_k \in \mathcal{V} \equiv \mathbb{R}^d\}_{k=1}^K$ sampled from a probability distribution P_v . StyleGANs [20] use a *Multilayer perceptron* (MLP) $F : \mathcal{V} \rightarrow \mathcal{W}$ that maps v_k to an intermediate latent space point w_k , which is then fed to a generator $G : \mathcal{W} \rightarrow \mathcal{I}$ that synthesizes an image $G(F(v_k))$. One trains the generator by pitting it against a discriminator network Q that distinguishes

between real and generated images. The disentangled latent space \mathcal{W} used in StyleGANs gives improved image generation compared to the basic GAN architecture. AE-StyleGAN2 [14] borrows ideas from a VAE, attaching an encoder E to the model which maps from the image to latent space (as our application requires) and giving additional training to G as a decoder. E and G are trained using autoencoder reconstruction loss $\min_{E,G} \|I - G(E(I))\| + \|\eta(I) - \eta(G(E(I)))\|$ where η is a pre-trained VGG16 network, and adversarial loss $\min_{E,F,G} \max_Q \mathbb{E}_I[\log Q(I)] + \mathbb{E}_v[\log(1 - Q(G(F(v))))] + \mathbb{E}_I[\log(1 - Q(G(E(I))))]$ which adds an autoencoder term to previous StyleGAN objectives. While AE-StyleGAN2 accomplishes its aims, it does not consider the geometry of the image manifold. *For preserving geometry, we propose GP-StyleGAN2.*

3.1 Learning Latent Map Using GP-StyleGAN2

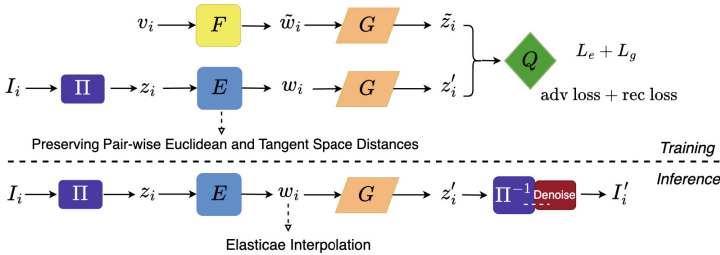


Fig. 1. Training and inference procedure. We preserve the geometry of image space by constraining the encoder (E) with loss functions L_e and L_g defined in Algorithms 1, 2.

GP-StyleGAN2 Architecture: We seek a **significant** dimension reduction (from $n^2 = 2^{14}$ to $d = 5$) which is (1) **invertible** and (2) **geometry preserving** ($d = 5$ is the smallest embedding dimension of $SO(3)$). The computational demands of learning a nonlinear map for such a drastic reduction suggest a two-step approach. We first move the problem to points $\{z_i\}$ in an intermediate dimension $c^2 = 2^{10}$ using PCA. This linear projection Φ suffices for a modest reduction while being approximately invertible and norm-preserving (Parseval’s theorem), though it will fail if we use it for the full reduction $2^{14} \rightarrow 5$. From here, we train the nonlinear encoder $E : \mathbb{R}^{c^2} \rightarrow \mathbb{R}^d$ generator $G : \mathbb{R}^d \rightarrow \mathbb{R}^{c^2}$ between $\{z_i\}$ in the PCA space and points $\{w_i\}$ in the latent space. After this training we build the finalized maps $\Phi : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^d$ and $\Phi^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^{n^2}$ defined by the compositions $\Phi = E \circ \Phi$ and $\Phi^{-1} = \Phi^{-1} \circ G$. We note that Φ^{-1} and Φ^{-1} are approximate inverses and that the PCA reconstruction Φ^{-1} is fine-tuned by a denoising neural network. The architecture for training E and G begins with the AE-StyleGAN2 autoencoder and adversarial objectives, but augments the training of E with geometry-preserving loss terms based on pairwise point and tangent space distances. Fig. 1 shows a schematic of the training and inference procedure.

Algorithm 1. Optimizing encoder E with geometry-preserving term L_e (points)

-
- 1: Given (1): E ; (2): A batch of images $I^B \in \mathbb{R}^{b \times n^2}$.
 - 2: Map images to PCA space: $Z^B = \Phi(I^B) \in \mathbb{R}^{b \times c^2}$.
 - 3: Map Z^B to latent space: $W^B = E(Z^B) \in \mathbb{R}^{b \times d}$.
 - 4: Compute pair-wise distance matrices: $D_{ij}^{e,z} = \|z_i - z_j\|$ and $D_{ij}^{e,w} = \|w_i - w_j\|$, where $z_i, z_j \in Z^B$, $w_i, w_j \in W^B$, and $\|\cdot\|$ denotes the Euclidean norm.
 - 5: Compute the loss: $L_e = L(D^{e,z}, D^{e,w})$, where L is defined in Eqn. 1.
 - 6: Update the weights: $\theta_E \leftarrow \text{ADAM}(\nabla_{\theta_E} L_e, \theta_E)$, where θ_E denotes the internal parameters of E , and ADAM refers to adaptive moment optimization [21].
-

Geometry-preserving Terms: The central feature of our method is a loss function designed to make the Euclidean distances between the latent space encodings of images correspond to their Euclidean distances in the image space. The loss is computed as a dissimilarity between pairs of distance matrices. We use three types of metrics to compute these matrices. The first is the standard Euclidean distance $\|\cdot\|$. The second is a metric defined on $SO(3)$: Using the rotation matrix representations $s_i, s_j \in SO(3)$ of two poses of an object, the Riemannian distance between them is $d_s(s_i, s_j) = \cos^{-1} \left(\frac{\text{trace}(s_i s_j^T) - 1}{2} \right)$. The third is a metric on sets of linear subspaces with common dimensions: Let $T_i, T_j \in \mathbb{R}^{d \times r}$, $r \leq d$, denote arbitrary orthogonal bases of any two r -dimensional subspaces in \mathbb{R}^d . Then, define $d_g(T_i, T_j) = \|T_i T_i^T - T_j T_j^T\|_F$. We use this extrinsic distance on the Grassmannian manifold to simplify computations.

In practice, we approximate tangent spaces in the PCA space and the latent space using the training data as follows. First, we find $N > 3$, $SO(3)$ -neighbors for each training image (s_i) using the lowest values of $d_s(s_i, s')$. For PCA points, we then approximate N tangent vectors at z_i using finite differences as $\{V_{ij} = \frac{z_j - z_i}{d_s(s_i, s_j)} \in \mathbb{R}^{c^2}\}_{j=1}^N$ and set $T_i^z \in \mathbb{R}^{c^2 \times 3}$ to be the three dominant singular vectors of the set $\{V_i\}$. Similarly, we can approximate tangent spaces $T_i^w \in \mathbb{R}^{d \times 3}$ in the latent space (see Algorithm 2). A very similar method for tangent plane estimation is used in [37], where a proof is given for convergence to the true tangent plane.

Given a batch of b images, we compute distance matrices $D \in \mathbb{R}^{b \times b}$ between mapped points and tangent planes in the PCA and latent spaces. Then a measure of discrepancy between computed matrices D^1 (PCA space) and D^2 (latent space) is calculated using:

$$L(D^1, D^2) = \sum_{j=1}^b \left[1 - \frac{(D_{\cdot j}^1 - \mu_j^1 \mathbf{1})^T (D_{\cdot j}^2 - \mu_j^2 \mathbf{1})}{\|D_{\cdot j}^1 - \mu_j^1 \mathbf{1}\| \|D_{\cdot j}^2 - \mu_j^2 \mathbf{1}\|} \right], \quad (1)$$

where μ_j^1 and μ_j^2 are the mean values for columns $D_{\cdot j}^1$ and $D_{\cdot j}^2$, and $\mathbf{1}$ is a vector of ones. We center and scale columns (or rows) of D matrices into unit vectors and compute the cosines of angles between them. In an implicit manner, each entry of D^1 is compared with the corresponding entry of D^2 . Since we are forming a loss function, we subtract this quantity from one, and sum over all points in the batch. We use the resulting loss L to define novel geometric terms for modifying AE-StyleGAN2 as follows:

1. **Term 1: Distance Preserving:** Here $D_{ij}^1 = \|z_i - z_j\|$, the Euclidean distances between PCA scores of training images, and $D_{ij}^2 = \|w_i - w_j\|$, the Euclidean distances between corresponding latent vectors. We use Euclidean distances between all pairs, not just the neighbors. These distances play the role of extrinsic or embedding distances between points on the (unknown) manifold and help learn its global geometry. Later on in the paper, once the manifold is estimated, we use geodesics and geodesic (intrinsic) distances to perform statistical analysis. In other words, we use the extrinsic Euclidean distance for learning and intrinsic geodesic distance for analysis. We will call the loss $L = L_e$ in this case. Algorithm 1 lists the steps for computing L_e .
2. **Term 2: Curvature Preserving:** Here $D_{ij}^1 = d_g(T_i^z, T_j^z)$, the tangent space distance, and $D_{ij}^2 = d_g(T_i^w, T_j^w)$ the tangent distances in the latent space. We will call the loss $L = L_g$ in this case. Algorithm 2 lists the steps for computing L_g .

Algorithm 2. Optimizing encoder E with geometry-preserving term L_g (tangent distances)

- 1: Given (1): A set of images $I^B \in \mathbb{R}^{b \times n^2}$. (2): The corresponding rotation set $S^B \in \mathbb{R}^{b \times 3 \times 3}$. (3): Corresponding neighborhoods $\mathcal{N}_i^I = [I_{\ell_1(i)}, \dots, I_{\ell_N(i)}] \in \mathbb{R}^{n^2 \times N}$, of the N closest points to $I_i \in I^B$ according to d_s . (4): The the index functions $\{\ell_k\}_{k=1}^N$, which take an argument i and return the index of the k th nearest neighbor of I_i .
 - 2: Map images and their corresponding neighbors to PCA space:
 $Z^B = \Phi(I^B) \in \mathbb{R}^{b \times c^2}$, $\mathcal{N}_i^z = \Phi(\mathcal{N}_i^I) = [z_{\ell_1(i)}, \dots, z_{\ell_N(i)}]$, where $z_i \in Z^B$.
 - 3: Map Z^B and $\{\mathcal{N}_i^z\}_{i=1}^b$ to latent space:
 $W^B = E(Z^B) \in \mathbb{R}^{b \times d}$, $\mathcal{N}_i^w = E(\mathcal{N}_i^z) = [w_{\ell_1(i)}, \dots, w_{\ell_N(i)}]$, where $w_i \in W^B$.
 - 4: Compute neighborhood $SO(3)$ distances:
 $\Delta_i^s = [d_s(s_i, s_{\ell_1(i)}) \cdots d_s(s_i, s_{\ell_N(i)})] \in \mathbb{R}^N$, where $s_i, s_j \in S^B$.
 - 5: Compute over-dimensional tangent planes in *Principal Component Analysis* (PCA) space and latent space:
 $\tilde{T}_i^z = (\mathcal{N}_i^z - z_i \mathbf{1}^T) \cdot \text{diag}(\Delta_i^s)^{-1} \in \mathbb{R}^{c^2 \times N}$, $\tilde{T}_i^w = (\mathcal{N}_i^w - w_i \mathbf{1}^T) \cdot \text{diag}(\Delta_i^s)^{-1} \in \mathbb{R}^{d \times N}$.
 - 6: Compute tangent planes T_i^z, T_i^w by taking the three dominant singular vectors of the corresponding $\tilde{T}_i^z, \tilde{T}_i^w$.
 - 7: Compute pair-wise distance matrices: $D_{ij}^{g,z} = d_g(T_i^z, T_j^z)$ and $D_{ij}^{g,w} = d_g(T_i^w, T_j^w)$.
 - 8: Compute the loss: $L_g = L(D^{g,z}, D^{g,w})$.
 - 9: Update the weights: $\theta_E \leftarrow \text{ADAM}(\nabla_{\theta_E} L_g, \theta_E)$.
-

3.2 Elasticae Interpolation

We can use the trained map Φ to project test data into the latent space. However, this data may be sparse, and we wish to discover the projected manifold \mathcal{M}^α at a higher resolution than the test data permits. One way to find intermediate points is through interpolation. Straight line interpolations would be reasonable if we had just the points $\{w_i \in \mathbb{R}^d\}$. However, our access to tangent planes $\{(w_i, T_i^w) \in \mathbb{R}^d \times \mathbb{R}^{d \times 3}\}$ at each

point allows us to account for the nonlinearity of the underlying manifold by utilizing nonlinear interpolations based on elasticae.

Elasticae are smooth curves that can be used to interpolate between directed points, *i.e.* Euclidean points with attached tangent vectors. Consider the set \mathcal{B} of smooth curves in \mathbb{R}^d parameterized on $[0, 1]$. For a curve $\beta \in \mathcal{B}$, let $\dot{\beta}$ and κ_β denote its velocity and scalar curvature functions, $Len[\beta]$ its length, and define its elastic energy $En[\beta] = \frac{1}{2} \int_0^1 \kappa_\beta^2(s) ds$. Then the *free elastica* from a given directed point (w_1, u_1) to another (w_2, u_2) is the minimizing curve $\hat{\beta} = \arg \min_{\beta \in \mathcal{B}} (En[\beta] + \lambda Len[\beta])$ such that $\beta(0) = w_1$, $\beta(1) = w_2$, $\dot{\beta}(0) = u_1$, and $\dot{\beta}(1) = u_2$. The tuning parameter $\lambda > 0$ balances the focus on length versus curvature. Mumford [29] advocated the use of free elasticae as the most likely solutions to fill in the missing or obscured curves in images, *e.g.*, in the famous Kanizsa triangles.

Our implementation follows Mio et al. ([28], Algorithm 4.2). We interpolate between point pairs w_i, w_j in the latent space \mathbb{R}^d . To find the corresponding tangent vectors, we take the difference vector $\tilde{u}_{ij} = w_j - w_i$, project it separately into each point’s tangent space, and scale to form the unit vectors $u_i \in \text{span}(T_i^w)$ and $u_j \in \text{span}(T_j^w)$. These maximally-aligned vectors are used to direct a free elastica β that interpolates from (w_i, u_i) to (w_j, u_j) . This interpolation is mapped to a path in image space as $\Phi^{-1}(\beta(t))$. The top portion of Fig. 3 in Section 4.1 illustrates the image space elasticae as sequences of images.

4 Experiments

Before detailing the design and results of our experiments, we begin by laying the groundwork of key features which are used throughout.

Experimental Data: Creating an image set to represent an object \mathcal{I}^α for training Φ and Φ^{-1} requires a structured sampling over rotations in $SO(3)$. We express rotations using a Hopf coordinate system similar to Yershova et al. [41]. Seeking a partition of $SO(3)$ with regular equivolumetric cells, we first generate (θ, ϕ) values on \mathbb{S}^2 using the Fibonacci grids of Swinbank and Purser [39] (also studied in Hardin et al. [15]). We attach a circle of ψ values to each pair, and the circle is uniformly segmented following a ratio shown in [41] to produce $SO(3)$ cells analogous to cubes. The experimental results presented here are restricted to a patch of $SO(3)$ to simplify computations. This patch is $\mathcal{P} \triangleq \{(\theta, \phi, \psi) \in [\frac{2}{5}\pi, \frac{3}{5}\pi] \times [\frac{4}{5}\pi, \frac{6}{5}\pi]^2\}$. To create an image set, we begin with a three-dimensional object O^α set at a default position. We use CAD objects $\alpha \in \{\text{chair, sports car, zebra}\}$ from clara.io [1] processed using meshio [34]. We apply 4000 rotations from \mathcal{P} to the default orientation and use \mathbb{P} to generate images $\mathbb{P}(s_i O^\alpha)$. We will call this training set $\mathcal{R}_{train} \triangleq \{(s_i, I_i) \in (SO(3) \times \mathcal{I}^\alpha)\}_{i=1}^{4000}$. Fig. 2 shows a representation of these points. We also create a separate data set \mathcal{R}_{test} of 3696 indexed images for testing and evaluation. The points in \mathcal{R}_{test} also lie in \mathcal{P} but are defined by a uniform rectangular grid in (θ, ϕ, ψ) coordinates.

Evaluation Metrics: Our goals include performing several tasks that can be evaluated quantitatively. In our quantifiable experiments, we compare model outcomes against

ground truth values using the Euclidean norm and tabulate errors. In our most-used scenario, we compare a ground truth path in image space $\{I(t) \in \mathbb{R}^{n^2}\}_{t=0}^T$ with an estimated $\{\hat{I}(t) = \Phi^{-1}(\hat{w}(t)) \in \mathbb{R}^{n^2}\}_{t=0}^T$ decoded from a computation in latent space. Our evaluations use the Squared Errors (SE) $\|\hat{I}(t) - I(t)\|^2$ indexed by t . For elasticiae evaluations, we also compare the velocities along the path to the ground truth using $\|(\hat{I}(t+1) - \hat{I}(t)) - (I(t+1) - I(t))\|^2$.

Model Comparisons: We compare our method with three recent deep-learning generative models described in Section 2: RHVAE [5], SVAE [8], and AE-StyleGAN2 [14].

Implementation details: Experiments are conducted on a Linux workstation with Nvidia RTX A6000 (48GB) GPU and Intel Core i9-13900K CPU @ 5.8GHz with 128GB RAM. The hyperparameters of generator G and MLP F are chosen to be identical to those in [19]. The hyperparameters of encoder E are the same as in [14]. The size of image space is $n^2 = 128^2$. The size of PCA space is $c^2 = 32^2$. The dimension of latent space is $d = 5$. The batch size for training is $b = 64$. The tangent spaces are built using $N = 16$ neighbors. The balance parameter for elasticiae is $\lambda = 1$.

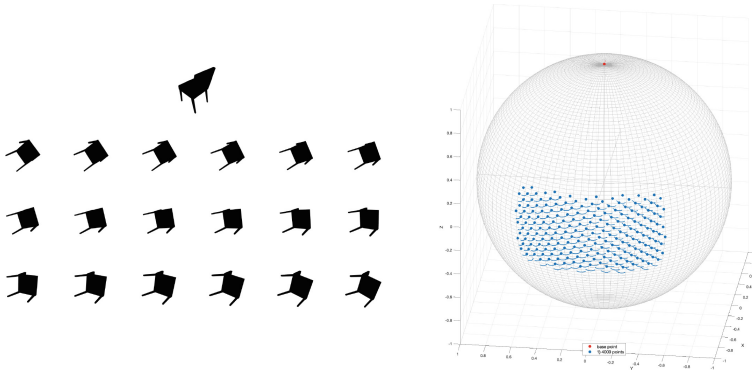


Fig. 2. Right: A visual representation of the 4000 $SO(3)$ training rotations. Each rotation is represented by its heading (θ, ϕ) on S^2 and its roll about that heading drawn as a curve ψ around it. Left: Top image is the chair at default orientation. Others are $\mathbb{P}(sI^{chair})$ for some $s \in SO(3)$ with the same heading and varying roll.

4.1 Results: Elasticiae Interpolations and Manifold Estimation

Having used \mathcal{R}_{train} to train the maps Φ and Φ^{-1} , we map elements of \mathcal{R}_{test} to points in latent space, use elasticiae to interpolate between them, and then map these elasticiae paths to image space. We then compare results from our framework to using various SOTA DNN models. Fig. 3 (top) shows results for $\alpha = chair$. These interpolations are computed in the latent space but visualized in the image space. Each row shows an interpolation between two fixed points in \mathcal{R}_{test} at the left and right. Different rows correspond to different techniques. The bottom row shows the ground truth geodesic $t \mapsto \mathbb{P}(x(t)O^\alpha)$, where $x(t)$ is a geodesic in $SO(3)$. We observe that the path obtained

by our model is consistently closest to the ground truth. This outcome is representative of the results we obtained for several experiments.

To quantify performance, we perform extensive experiments on three 3D objects: $\alpha \in \{\text{chair, sports car, zebra}\}$. We compute 100 different interpolation paths using randomly selected pose pairs in \mathcal{R}_{test} (3696 points). For each time index t , we calculate the mean values of the errors (point values and tangent values) and plot them in Fig. 3. The errors are naturally close to zero at the start and the end, and are highest at the center. As these plots exhibit, the interpolation errors are the smallest using our method when compared to RHVAE, SVAE, and AE-StyleGAN2.

Ablation Studies: To evaluate the key components of GP-StyleGAN2, we perform ablation studies that add them sequentially to an AE-StyleGAN2 baseline model. We study six models which differ in: (1) the type of interpolation: linear or elastica, (2) the loss function for training E : inclusion of geometry-preserving terms L_e and L_g (defined in Section 3.1) or not, and (3) PCA for image pre-processing: PCA or no PCA. For each of the six models, the experimental setup is an analysis of interpolation path accuracy as above. The results of these experiments performed on the chair object are summarized in Table 1. The table lists the means over 100 experiments of the average interpolation error summed over all time indices t in the path. Comparing Models 1, 2 versus the others shows dramatic gains due to the introduction of the geometry-preserving terms. Comparisons of Models 3, 4, and 5 versus Model 6 show individual benefits of the use of both L_e and L_g over L_e alone, elasticae over linear interpolation, and PCA reduction.

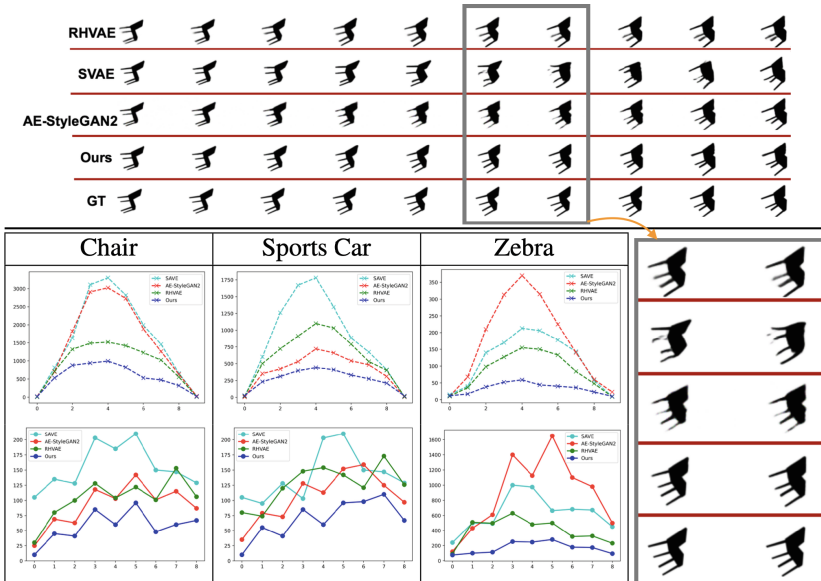


Fig. 3. Top: Interpolated paths between the original (left) and rotated poses (right) using various methods. Bottom: Average squared errors over 100 interpolated paths for each object, and zoom from top. First row: Average SEs for interpolated points. Second row: Average SEs for velocities.

The complete Model 6 (GP-StyleGAN2) including all of the components substantially outperforms the others.

Computational Cost: The computational cost of training for the four methods are as follows: AE-StyleGAN2 - 20.64 *min*/500 epochs, RHVAE - 12.78 *min*/500 epochs, SVAE - 8.78 *min*/500 epochs, and GP-StyleGAN2 - 38.41 *min*/500 epochs. The computational cost for interpolating a path between two test points is: AE-StyleGAN2 - 0.02 sec, RHVAE - 62.36 sec, SVAE - 0.01 sec, and GP-StyleGAN2 - 0.13 sec.

Manifold Estimation: To estimate the latent pose manifold $\widehat{\mathcal{M}}^\alpha$, we randomly select 800 sparse points from \mathcal{R}_{test} and map them using the trained Φ . For each point in this latent space, we identify its five nearest-neighbors using $SO(3)$ distance d_s and interpolate between these neighbors using elasticae with eight intermediate points. This results in an $\widehat{\mathcal{M}}^\alpha$ with 32,800 total points.

Table 1. (Ablation studies): Average total squared errors over 100 interpolated paths (for chair) under different models. AE-loss denotes the standard loss function of AE-StyleGAN2.

Model Features	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AE-StyleGAN2	yes	yes	yes	yes	yes	yes
Interpolation	linear	elasticae	elasticae	linear	elasticae	elasticae
Loss function	AE-loss	AE-loss	AE-loss + L_e	AE-loss + $L_e + L_g$	AE-loss + $L_e + L_g$	AE-loss + $L_e + L_g$
PCA	no	no	yes	yes	no	yes
Summary	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 (Ours)
Mean error	1638.97	1654.32	698.14	687.02	561.70	436.11

4.2 Results: Analyzing Latent Map Φ Using Test Data

We investigate the geometry-preserving properties of the mapping Φ by applying it to the complete test set \mathcal{R}_{test} . As described above, \mathcal{R}_{test} is a set of 3696 points in the same patch \mathcal{P} as \mathcal{R}_{train} but disjoint from it. We exploit the grid structure of \mathcal{R}_{test} to interpret the range space of Φ visually, create paths that bridge distant points, and verify the goal of distance preservation on a large scale.

Visualizing $\Phi(\mathcal{R}_{test})$: First we visualize the mapping of \mathcal{R}_{test} into \mathbb{R}^d using the learnt Φ . Here we investigate the geometry of the underlying manifold \mathcal{M}^α using this set’s grid structure rather than densifying interpolations. The latent space plots shown here use the first three PCA axes of the mapped points in \mathbb{R}^5 . We note that the subset of $SO(3)$ used in these experiments is topologically a box and that the first three singular values accounted for most of the variation in these examples. Figure 4 displays two viewing angles of the GP-StyleGAN2 latent space embeddings of the chair, car, and zebra objects. The three clouds look remarkably similar, all resembling shell-like segments of a thickened sphere, despite the vastly different shapes of the original objects.

Traversing Distant Points in $\Phi(\mathcal{R}_{test})$: In this experiment, we first endow the set $\Phi(\mathcal{R}_{test})$ with a graph structure determined by $SO(3)$ neighbors. Each point in $SO(3)$ has 26 neighbors; see the supplement for details. We then arbitrarily select two distant points in this set and create three paths between them: (1) the ground truth path

derived from the geodesic in $SO(3)$, (2) the shortest-length path through the graph found using Dijkstra’s algorithm, and (3) a simple straight-line interpolation in latent space. Finally, we map the paths to image space. Figure 5 compares the results obtained using AE-StyleGAN2 and GP-StyleGAN2. We can derive multiple conclusions from these results. Firstly, the linear interpolations perform poorly under both models, highlighting the nonlinearity of the pose manifold. Notice how the linear interpolation loses its chair structure as it passes through the hollow space in the point cloud on the right. Secondly, GP-StyleGAN2 performs significantly better than AE-StyleGAN2 both visually and by error quantification: the Dijkstra path for GP-StyleGAN is nearly as good as the ground truth, while for AE-StyleGAN it is hardly any better than the straight line.

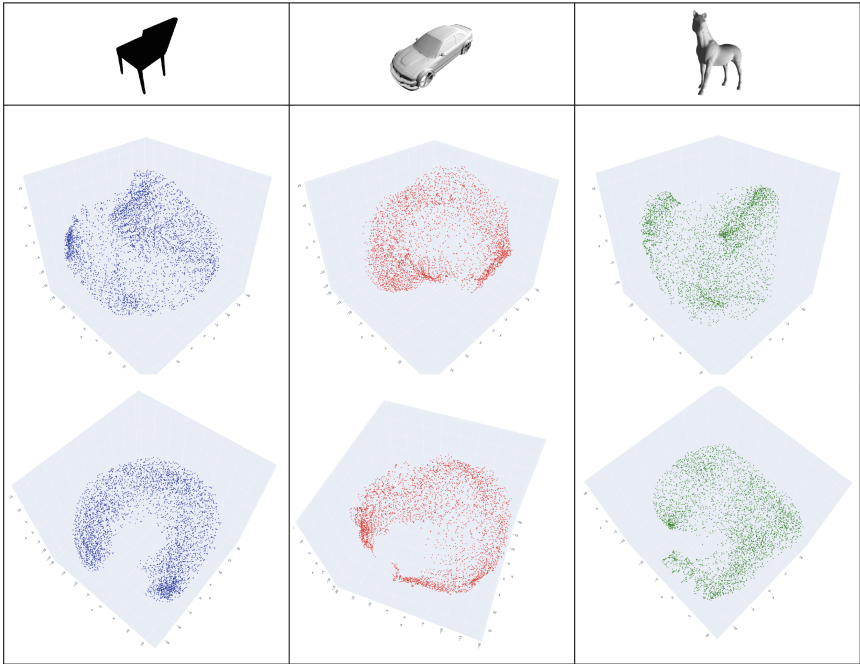


Fig. 4. Latent pose manifolds of chair, car, and zebra objects. Two views of each. We emphasize that these are not just points but are graphs with geometries.

4.3 Results: Exploiting Manifold Geometry

Once we have the estimated manifold $\widehat{\mathcal{M}}^\alpha$, we study its geometry in two different ways: Computing intrinsic image means and performing image denoising. (Fig. 6).

Mean Computations on $\widehat{\mathcal{M}}^\alpha$: We assess the utility of finding mean images in the latent space by comparing the decoded means to the ground truth, defined by images associated with $SO(3)$ Karcher means of sample rotations. Selecting 10 random images

$\{I_j \in \mathcal{R}_{test}\}$, we first compute a naive mean in \mathbb{R}^{n^2} , denoted as μ_I . Then we compute the Euclidean mean of $\{\Phi(I_j)\}$ in \mathbb{R}^d , project it to the nearest point in $\widehat{\mathcal{M}}^\alpha$, and map it back to image space to define μ_w . For comparison, we show the result $\tilde{\mu}_w$ obtained by performing these steps but skipping projection. The corresponding quantities under AE-StyleGAN2 are labeled as μ_w^A and $\tilde{\mu}_w^A$ respectively. Fig. 6 (top part) compares these results with the ground truth means μ_{gt} . The means estimated using GP-StyleGAN2 (green boxes) display realistic structure of the chair and better resemble the ground truth (orange boxes) than AE-StyleGAN2.

Image Denoising using $\widehat{\mathcal{M}}^\alpha$: The manifold geometry can also be used for denoising or cleaning corrupted images. A noisy image J can be mapped into latent space a $\Phi(J)$, projected to the nearest point $w \in \widehat{\mathcal{M}}^\alpha$, and mapped back as a cleaned image $\Phi^{-1}(w)$. Fig. 6 (bottom) shows images of the chair corrupted by adding noise and clutter, and compares results of cleaning using GP-StyleGAN2 and AE-StyleGAN2. The visual results and a histogram of reconstruction errors both show better outcomes using GP-StyleGAN2.

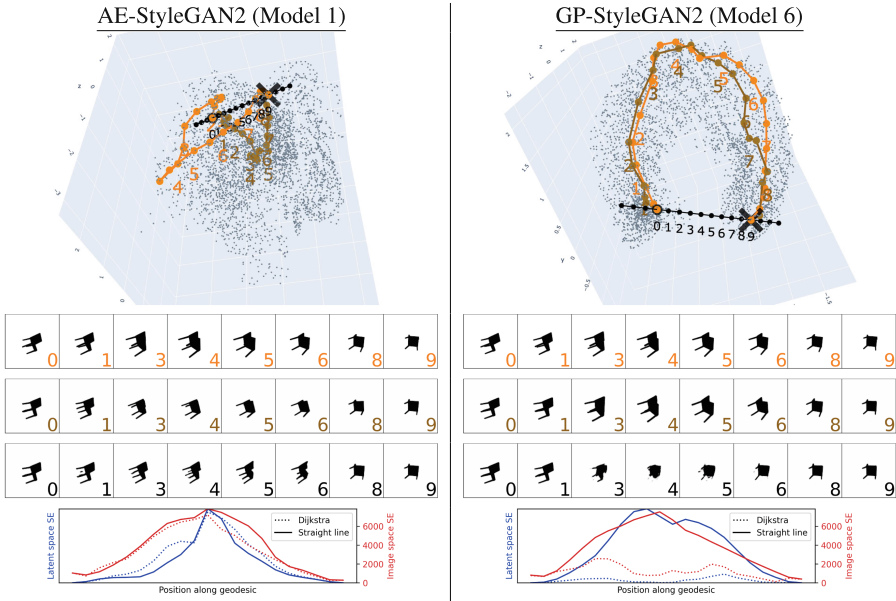


Fig. 5. Traversing distant points. Top and middle rows: Paths in latent and image space - geodesic GT (orange), Dijkstra on graph (brown), and straight line (black). Bottom: Euclidean squared errors along paths. Dijkstra paths under GP-StyleGAN2 perform best.

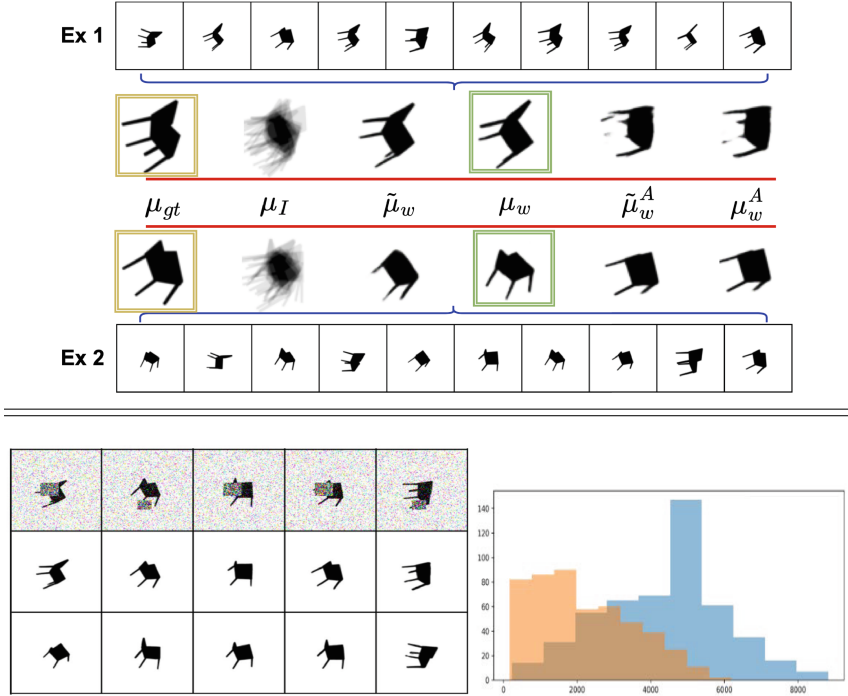


Fig. 6. Top Part: Manifold averaging on $\widehat{\mathcal{M}}^\alpha$. Ground truth μ_{gt} is compared with image space mean μ_I and different latent space means: $\mu_w, \tilde{\mu}_w$ computed with/without projection on $\widehat{\mathcal{M}}^\alpha$ using GP-StyleGAN2, and likewise $\mu_w^A, \tilde{\mu}_w^A$ using AE-StyleGAN2. Bottom Part: Left: Top row displays noisy images J , middle and bottom rows show corresponding denoised images using GP-StyleGAN2 and AE-StyleGAN2, respectively. Right: Histogram of squared errors for 500 noisy images using AE-StyleGAN2 (blue) and GP-StyleGAN2 (orange).

5 Conclusions

We introduced a new approach, GP-StyleGAN2, for characterizing pose manifolds of 3D objects. This approach preserves geometry when mapping to a low-dimensional latent space and creates dense manifold representations that account for nonlinearity using Euler’s free elasticae. Comparisons of interpolations using GP-StyleGAN2 and various other methods (Fig. 3) showed superior results for our model visually and quantitatively. Ablation studies (Table 1) gave more detailed quantitative results that demonstrated improvements from including our two novel geometry-preserving terms and using elasticae rather than linear interpolation. Graph-based geodesic approximations pointed to regular but nonlinear geometry of pose manifolds (Fig. 5), in stark contrast to past conclusions of linear geometry for latent space image data. We also found that the use of manifold geometry improved mean computations and image denoising (Fig. 6).

While there is still much progress to be made, the overall success of GP-StyleGAN2 shows a step in the direction of truly learning the geometry of pose manifolds.

Acknowledgements. This research was supported in part by the grants DARPA-PA-21-04-05-FP-052, NIH R01-GM135927, and NSF IIS 1955154 to AS and NSF award 2323086 to PT.

References

1. Clara.io. <https://clara.io>
2. Arvanitidis, G., Hansen, L.K., Hauberg, S.: Latent space oddity on the curvature of deep generative models. In: Proc. of International Conference on Learning Representations (2018)
3. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013)
5. Chadebec, C., Allasonniere, S.: A geometric perspective on variational autoencoders. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) *Advances in Neural Information Processing Systems* (2022), <https://openreview.net/forum?id=PBmJC6rDnR6>
6. Chen, N., Klushyn, A., Ferroni, F., Bayer, J., Van Der Smagt, P.: Learning flat latent manifolds with vaes. arXiv preprint [arXiv:2002.04881](https://arxiv.org/abs/2002.04881) (2020)
7. Coifman, R.R., Lafon, S.: Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**(1), 5–30 (2006)
8. Davidson, T.R., Falorsi, L., De Cao, N., Kipf, T., Tomczak, J.M.: Hyperspherical variational auto-encoders. arXiv preprint [arXiv:1804.00891](https://arxiv.org/abs/1804.00891) (2018)
9. Detlefsen, N.S., Hauberg, S., Boomsma, W.: Learning meaningful representations of protein sequences. *Nat. Commun.* **13**(1), 1914 (2022)
10. Donoho, D.L., Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. of the National Academy of Sciences* **100**(10), 5591–5596 (2003)
11. Duque, A.F., Morin, S., Wolf, G., Moon, K.R.: Geometry regularized autoencoders. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(6), 7381–7394 (2022)
12. Goodfellow, I., et al.: Generative adversarial networks. *Comm. of the ACM* **63**(11), 139–144 (2020)
13. Grenander, U., Srivastava, A., Miller, M.I.: Asymptotic performance analysis of Bayesian object recognition. *IEEE Trans. Inf. Theory* **46**(4), 1658–66 (2000)
14. Han, L., Musunuri, S.H., Min, M.R., Gao, R., Tian, Y., Metaxas, D.: AE-StyleGAN: Improved training of style-based auto-encoders. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 3134–3143 (2022)
15. Hardin, D.P., Michaels, T., Saff, E.B.: A comparison of popular point configurations on \mathbb{S}^2 . arXiv preprint [arXiv:1607.04590](https://arxiv.org/abs/1607.04590) (2016)
16. Hauberg, S.: Only Bayes should learn a manifold (on the estimation of differential geometric structure from data). arXiv preprint [arXiv:1806.04994](https://arxiv.org/abs/1806.04994) (2018)
17. Huh, I., et al.: Isometric quotient variational auto-encoders for structure-preserving representation learning. *Neural Information Processing Systems* **36** (2024)
18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: *International Conference on Learning Representations* (2018)
19. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. *Adv. Neural. Inf. Process. Syst.* **33**, 12104–12114 (2020)

20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Computer Vision and Pattern Recognition*. pp. 4401–4410 (2019)
21. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*. San Diego, CA, USA (2015)
22. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) (2013)
23. Kühnel, L., Fletcher, T., Joshi, S., Sommer, S.: Latent space geometric statistics. In: *Pattern Recognition. ICPR International Workshops and Challenges*, pp. 163–178. Springer International Publishing, Cham (2021)
24. Lim, J., Kim, J., Lee, Y., Jang, C., Park, F.C.: Graph geometry-preserving autoencoders. In: *Forty-first International Conference on Machine Learning* (2024)
25. Linnér, A.: Existence of free nonclosed euler-bernoulli elastica. *Nonlinear Analysis: Theory, Methods & Applications* **21**(8), 575–593 (1993)
26. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *Journal of machine learning research* **9**(11) (2008)
27. McQueen, J., Meila, M., Joncas, D.: Nearly isometric embedding by relaxation. *Advances in Neural Information Processing Systems* **29** (2016)
28. Mio, W., Srivastava, A., Klassen, E.: Interpolations with elasticae in euclidean spaces. *Q. Appl. Math.* **62**(2), 359–378 (2004)
29. Mumford, D.: *Elastica and computer vision* p. 491–506 (1994)
30. Sáez de Ocáriz Borde, H., Arroyo, A., Morales, I., Posner, I., Dong, X.: Neural latent geometry search: Product manifold inference via gromov-hausdorff-informed bayesian optimization. *Advances in Neural Information Processing Systems* **36** (2024)
31. Pai, G., Talmon, R., Bronstein, A., Kimmel, R.: Dimal: Deep isometric manifold learning using sparse geodesic sampling. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp. 819–828. IEEE (2019)
32. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434) (2015)
33. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
34. Schlömer, N.: meshio: Tools for mesh files. <https://doi.org/10.5281/zenodo.1173115>, <https://github.com/nschloe/meshio>
35. Shao, H., Kumar, A., Fletcher, P.T.: The Riemannian geometry of deep generative models. arXiv [abs/1711.08014](https://arxiv.org/abs/1711.08014) (2017)
36. Shukla, A., Uppal, S., Bhagat, S., Anand, S., Turaga, P.: Geometry of deep generative models for disentangled representations. *ICVGIP 2018, Association for Computing Machinery, New York, NY, USA* (2020)
37. Singer, A., Wu, H.T.: Vector diffusion maps and the connection laplacian. *Commun. Pure Appl. Math.* **65**(8), 1067–1144 (2012)
38. Singh, A., Nag, K.: Structure-preserving deep autoencoder-based dimensionality reduction for data visualization. In: *2021 IEEE/ACIS 22nd International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. pp. 43–48. IEEE (2021)
39. Swinbank, R., James Purser, R.: Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* **132**(619), 1769–1793 (2006)
40. Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *science* **290**(5500), 2319–2323 (2000)
41. Yershova, A., Jain, S., Lavalle, S.M., Mitchell, J.C.: Generating uniform incremental grids on SO(3) using the hopf fibration. *Intl. journal of robotics research* **29**(7), 801–812 (2010)

42. Yonghyeon, L., Yoon, S., Son, M., Park, F.C.: Regularized autoencoders for isometric representation learning. In: International Conference on Learning Representations (2021)
43. Zhang, S., Jiang, W.: Data-informed geometric space selection. *Advances in Neural Information Processing Systems* **36** (2024)
44. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.* **26**(1), 313–338 (2004)



MOMA: Contrastive Learning Distills Better Masked Autoencoders

Yuchong Yao^(✉), Nandakishor Desai, and Marimuthu Palaniswami

Department of Electrical and Electronic Engineering, University of Melbourne,
Parkville, VIC 3052, Australia
yuchongy1@student.unimelb.edu.au,
{nandakishor.desai,palani}@unimelb.edu.au

Abstract. Self-supervised learning has achieved remarkable performance in computer vision, utilizing two key paradigms: contrastive learning and masked image modeling. Contrastive learning focuses on global representations by learning similarities and dissimilarities from different views of the inputs. On the other hand, masked image modeling learns from a pixel-level reconstruction objective and has shown improved performance compared to contrastive learning. However, masked image modeling lacks global semantics due to its pixel-level objective. To this end, we propose MOMA, a novel self-supervised distillation framework that employs a contrastive learning teacher to enhance the global representation of the masked image modeling student. Specifically, the teacher provides masks for the student, encouraging reconstructions that favor better global semantics. The feature alignment between the teacher and the student further enhances the global features in masked image modeling. Experimental results demonstrate that the proposed MOMA outperforms other masked image modeling methods and achieves competitive performance compared to other self-supervised baselines.

Keywords: Self-supervised Learning · Knowledge Distillation · Computer Vision · Machine Learning · Deep Learning

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful methodology in various vision tasks and applications, particularly in computer vision [20, 21]. SSL eliminates the need for dataset annotations, reducing the costs associated with manual labeling and expertise. By extracting semantically rich knowledge from large volumes of unlabeled data, SSL forms the basis for powerful and generalizable models [3, 14]. The learned representations from these models can be effectively utilized in downstream tasks, and in some cases, they can even surpass the performance of supervised approaches. Among the rapidly evolving SSL methodologies in computer vision, two branches have established their dominance: contrastive learning and masked image modeling. Contrastive learning [7, 21] enhances unsupervised learning by emphasizing the agreement between two distinct augmented

views of the same input and enforcing disagreement with augmented views from different inputs. The key to this approach lies in applying reliable and challenging data augmentations to foster semantically significant representations. It learns the similarity or dissimilarity between the augmented views of the data, equipping the model with rich representation and high-level global semantics [7, 35]. Over the years, contrastive learning has showcased unprecedented results, even outperforming supervised learning algorithms in some cases [11]. However, it relies heavily on data augmentation [19] during its self-supervised pre-training and also requires a large batch size [8, 31] to possess adequate negative samples for the contrastive objective [29]. Recently, masked image modeling [20, 37] has emerged as another primary paradigm in self-supervised learning. Inspired by the success of masked language pre-training [4, 14] in natural language processing, masked image modeling aims to reconstruct original images from partially masked inputs. The framework adopts an encoder-decoder architecture, where the encoder encodes the masked inputs, and the decoder reconstructs the original inputs. The learning objective is to minimize the reconstruction loss in the pixel space. Notably, masked image modeling demonstrates high efficiency under high mask ratios, without needing hand-crafted data augmentation and large batch sizes, outshining contrastive learning across various benchmarks [13, 24]. However, the limitation of masked image modeling lies in its pixel-level reconstruction objective. Although this objective is simple and effective, it cannot capture high-level semantics and global features from the data [39, 41], as masked image modeling aims solely to reconstruct the masked pixels. Additionally, the masking process in masked image modeling is typically performed by random masking, which lacks semantics related to the global and discriminative features in the input. We question *whether we can encourage global representation learning in masked image modeling, thus fostering better representation and a more powerful self-supervised learning framework.*

To this end, we propose MOMA, a self-supervised learning framework incorporating knowledge distillation to encourage global feature learning and improve masked image modeling. MOMA forms a self-supervised distillation setup with an off-the-shelf contrastive learning teacher [11] and a masked image modeling pipeline as the student. The teacher provides two types of guidance for the student: (1) it first presents the attention map to guide the masking process for the masked image modeling pipeline, and (2) it also provides the target representations for the student encoder to align its learned features. The masks generated from the attention map encourage the student to reconstruct the most discriminative features and global semantics in the input. Furthermore, global semantics are strengthened during the feature alignment between the teacher and the student encoder. In this setup, the contrastive teacher effectively distills global features and high-level semantics to the masked image modeling student through a semantic-guided masking strategy and feature alignment. The major contributions of MOMA can be summarized as follows:

- We propose MOMA, a novel self-supervised knowledge distillation framework that effectively encourages global semantics and features for masked image modeling so that learning does not solely depend on pixel-level reconstruction.
- We utilize an off-the-shelf contrastive learning teacher to provide an attention map that guides the student’s masked image modeling pipeline to reconstruct features corresponding to the discriminative information and global semantics.
- We further align the learned features in the student encoder with the teacher through feature alignment, thereby encouraging the global semantics and features in the student’s representation.

Experimental results show that the proposed MOMA improves the performance of masked image modeling approaches, achieving competitive performance over various self-supervised baselines.

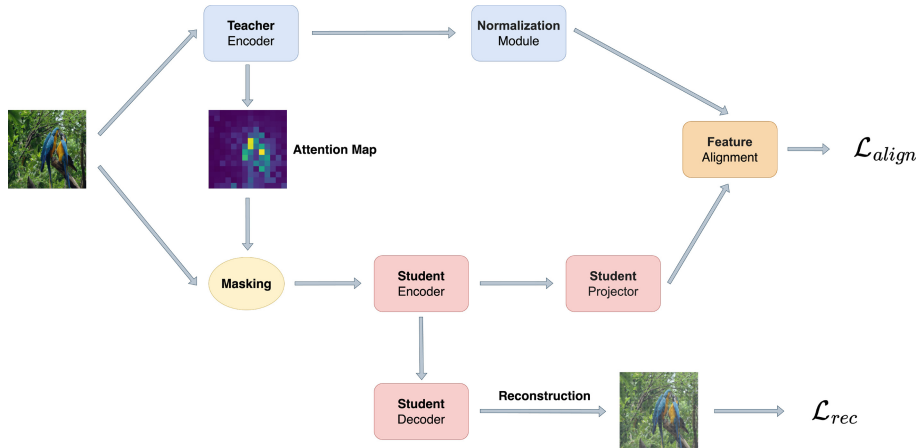


Fig. 1. Overview of the Proposed MOMA Framework. Our proposed self-supervised distillation framework consists of a static teacher branch and an actively updated student branch. The teacher, a pre-trained contrastive learning model, generates an attention map that guides the masking process in the student’s Vision Transformer encoder. The student representations are aligned with the teacher’s to reinforce semantics and global features. The framework optimizes for both a reconstruction loss (\mathcal{L}_{rec}) from masked image modeling and an alignment loss (\mathcal{L}_{align}) from feature alignment.

2 Related Work

2.1 Contrastive Learning

This self-supervised learning approach is based on instance discrimination [35], where each data sample is treated as an individual class. Each instance undergoes

substantial data augmentation, with positive pairs stemming from augmented views of the same instance and negative pairs from different instances. The learning process hinges on amplifying the concordance among positive samples (or disagreement between negative pairs), formally addressed by the InfoNCE loss [29]. SimCLR [7] and MoCo [21] are two of the most influential works that significantly advance contrastive learning, even showing better performance than supervised methods. SimCLR highlights the importance of projectors in the contrastive learning framework, while MoCo proposes a momentum encoder to further improve performance. In contrastive learning, having an adequate number of negative samples is critical so that the learning algorithm will not yield trivial solutions, as demonstrated by the improved performance of SimCLR implementation with large batch size [8]. Data augmentation also plays a key role in contrastive learning [8, 19], ensuring that the contrastive objective has diverse augmented views for quality representation learning. DINO [5] incorporates self-distillation into the contrastive learning framework and utilizes the Vision Transformer (ViT) [15], which shows exceptional performance, with the resulting attention maps of the self-supervised model very close to the ground-truth segmentation masks on the images. MoCo v3 [11] further improves the contrastive learning baselines by incorporating advanced training strategies, large batch sizes, and more advanced architectures, and stabilizes the training by freezing the patch embedding of ViT. Although contrastive learning methods achieve exceptional performance in various tasks in computer vision, their performance is limited by high reliance on large batches, hand-crafted heavy data augmentation, quality of negative samples, and strategies to ensure training stability. These requirements and constraints limit the usage and applicability of contrastive learning in different domains and motivate more straightforward self-supervised objectives.

2.2 Masked Image Modeling

This simple idea involves reconstructing corrupted input to form the self-supervised learning objective. The pioneering work [30] introduced inpainting as a pretext task for self-supervised learning, enabling the reconstruction of corrupted inputs. iGPT [6] performed reconstruction on corrupted images, adhering to the auto-regressive approach detailed in GPT [4]. Conversely, BEiT [3] adopted a BERT [14]-style pre-training protocol, which restores masked image tokens in an autoencoding fashion. The approach employed a pre-trained tokenizer to transform input images into visual tokens. MAE [20] and SimMIM [37] are two concurrent influential works that employ an end-to-end framework with an asymmetric encoder-decoder architecture, utilizing a high mask ratio to boost computational efficiency and challenge the model to learn better representations. The straightforward concept involves a masking strategy that can be as simple as random masking. In MAE, the encoder takes unmasked patches, and the decoder reconstructs the original images based on encoded visible tokens and masked tokens, outperforming previous contrastive learning benchmarks. SimMIM encodes both visible patches and the masks, supporting both vision

transformers [15] and hierarchical vision transformers (Swin [25]). CIM [16] utilized a generator to produce corrupted patches onto the original input rather than using masks, achieving competitive results on various vision benchmarks using both ViT and Convolutional Neural Networks (CNN). CAE [10] introduced a regressor component into the masked image modeling framework and formulates masked prediction and reconstruction objectives, which shows better transfer performance over classification and segmentation tasks. Although masked image modeling appears as a new paradigm in self-supervised learning and achieves improved performance than contrastive learning baselines, it lacks high-level semantics and global features as the learning objective is formed at the pixel level.

2.3 Knowledge Distillation in Self-supervised Learning

The concept of knowledge distillation was introduced in [22] and represents a technique to transfer knowledge from a well-trained teacher model to a more compact or compressed student model. Existing methods have also incorporated knowledge distillation into the self-supervised learning framework to improve the original baselines for various objectives. In [28], the authors used knowledge distillation to decouple the backbone model used for self-supervised pre-training and the supervised downstream tasks, creating a more flexible self-supervised framework with improved results. SEED [17] applies self-supervised distillation to enable contrastive learning for small models. S2-BNN utilized knowledge distillation to construct binary neural networks [12] from contrastive learning-based real-valued models. DMAE [2] demonstrated masked knowledge distillation on the intermediate features between the student and teacher for masked image modeling, where the teacher is a large pre-trained MAE encoder model, and the student is a standard MAE pipeline with a smaller encoder. The authors show that their proposed method can distill a student with comparable performance to the teacher under an extremely high mask ratio.

2.4 Combining Self-supervised Paradigms

Recent works have sought to fuse the strengths of contrastive learning and masked image modeling to compensate for individual limitations and encourage better representation learning. SiameseIM [33] incorporated masking as a part of data augmentation operations into the contrastive learning framework. The results show that SiameseIM improves the performance of classification and segmentation, with more significant improvement in few-shot learning and robustness. MimCo [41] strived to enhance the linear separability of masked image modeling by introducing a two-stage pre-training process that includes contrastive learning and masked image modeling. It achieves exceptional performance on small ViT models and outperforms other self-supervised baselines. CAN [27] applied a mask to both branches in a siamese network and optimized an InfoNCE loss [29], a reconstruction loss, and a denoising loss. The combination results in a simple and robust self-supervised learning algorithm that

outperforms methods relying solely on contrastive learning or masked image modeling.

3 Methodology

3.1 Preliminary

Momentum Contrast (MOCO). The proposed MOMA utilizes a pre-trained contrastive learning model (MoCo v3 [11]) as the teacher in the self-supervised distillation framework. MoCo v3 utilizes a siamese network setup with one main encoder and one momentum encoder, optimizing its contrastive learning objective as follows:

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^K \exp(q \cdot k_i / \tau)} \quad (1)$$

Here, q is the encoded query from the main encoder and k is the encoded key from the other branch’s momentum encoder (k_+ is the positive key), and τ is the temperature parameter for the contrastive objective [35]. The positive and negative keys are generated from strong data augmentations [8, 19] to create diverse views. Therefore, MoCo v3 enables the model to learn rich global information that captures the high-level discriminative features of the object that is invariant and robust to different augmented views. We take the main encoder from the pre-trained MoCo v3 framework as the off-the-shelf teacher network to encourage high-level semantic and global feature learning for the student in the masked image modeling pipeline.

Masked AutoEncoder (MAE). The Masked Autoencoder (MAE) [20] is built based on the simple design of pixel reconstruction from randomly masked inputs and trains the network end-to-end. It adopts an asymmetric encoder-decoder architecture, where the encoder learns the rich semantics from the data and the lightweight decoder performs the pixel reconstruction from the masked inputs. The pixel-level reconstruction objective can be formally addressed as follows

$$\mathcal{L}_{rec} = \mathcal{L}_2(\mathcal{D}_\theta, \mathcal{E}_\theta(\mathbf{x} \odot \mathcal{M}), \mathbf{x}) \quad (2)$$

Here, \mathcal{E}_θ and \mathcal{D}_θ are the encoder and decoder in MAE, respectively. \mathcal{M} stands for the mask applied on the input x . \odot is the operator that uses indices from \mathcal{M} to mask the input. In the proposed MOMA, we adapt MAE into the student branch, where the encoder in the adapted MAE is the student network in the self-supervised distillation network.

3.2 MOMA

Self-supervised Distillation Framework. We propose a self-supervised distillation framework to transfer high-level semantics and global features from a

contrastive learning-based teacher to a masked image modeling student. The proposed framework employs a Siamese network structure comprising two branches: a teacher branch and a student branch. In the teacher branch, we utilize an off-the-shelf MoCo v3 pre-trained Vision Transformer (ViT) as the teacher network, which receives unmasked input data. Once the inputs are encoded by the teacher network, they proceed to a normalization module that produces normalized features, serving as the target representations for the student network to align its learned representations. Layer normalization [1] is applied within the normalization module to stabilize and generalize the model [38]. Furthermore, the teacher network conveys its attention map to the student model to guide the masking process in the masked image modeling pipeline. The attention map, derived from the contrastive learning teacher, contains rich high-level semantics, global and discriminative features, thereby guiding the masks to foster improved learning of global features and enhance the discriminative power of the student. The student branch adopts a masked image modeling pipeline, consisting of three components: an encoder, a decoder, and a projector. The encoder is a ViT, the decoder is a shallow transformer with two layers, and the projector is a lightweight two-layer Multi-Layer Perceptron (MLP). The student branch receives masked input data, where the teacher network influences the masking process. The encoder processes the masked data, and the resulting encoded representations are passed to the decoder and the projector, respectively. The decoder aims to reconstruct the original input data, fulfilling the masked image modeling objective. Meanwhile, the projector maps the encoded representations from the encoder, aligning these projected representations with the target representations produced by the teacher branch. An overview of the proposed framework is illustrated in Fig. 1.

Masking Strategy. Although random masking suggested in MAE [20] employs a high mask ratio (i.e., 75%), it does not inherently prompt the model to focus on global semantics or discriminative features. To address this, we designed a semantic-guided masking strategy utilizing the attention map from the contrastive learning teacher. Specifically, the attention output from the last transformer block of the ViT in the contrastive learning teacher is extracted, retrieving the multi-head attention of the [CLS] token to all other tokens, excluding itself. The attention values are then averaged across the head dimension and reshaped to match the input image dimensions. Sorting the attention values in descending order, we identified that larger values correspond to more attended features, thus indicative of discriminative features that capture global semantics. The indices of the top 50% attention values are retained as the mask indices for the masking process. This approach encourages the masked image modeling pipeline to learn and reconstruct the most significant discriminative semantic features that encapsulate the global semantics of the objects.

Feature Alignment. To enforce the high-level semantics and global features, we implement feature alignment between the representations from the teacher and student branches. During pre-training, the student updates its parameters

not only to fulfill the masked image modeling objective but also to align its learned representation with the target representations from the teacher, thereby enhancing the global features. We align the normalized representations from the teacher branch with the projected representations from the student branch by minimizing a distance metric. For more stable and robust learning, we employ the smooth L_1 loss as the distance metric [18], which encourages better alignment of the teacher and student representations. The formalization of the alignment process is expressed as:

$$\mathcal{L}_{align} = \mathcal{L}_{SmoothL_1}(\text{Norm}(z_t), \text{Proj}(z_s)) \quad (3)$$

where Norm and Proj represent the normalization module and projector in the teacher and student branches, respectively. z_t and z_s denote the representations from the contrastive learning teacher and the student encoder, respectively.

Learning Objectives. The proposed self-supervised distillation framework jointly optimizes two objectives (see Eq. 4): a reconstruction loss (see Eq. 2) from the masked image modeling, where the masked indices are guided by the teacher network rather than random masking, and an alignment loss (see Eq. 3) between the representations from the two branches.

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{align} \quad (4)$$

Parameters of the teacher branch remain fixed during the self-supervised pre-training phase. The off-the-shelf ViT is frozen with no gradient updates, and the normalization module (without affine transformations) has no learnable parameters. In contrast, the student branch is dynamic, with its encoder, decoder, and projector actively updated through gradient backpropagation.

4 Experiments

4.1 Datasets and Experiment Setup

Datasets. During the pre-training stage, we utilized ImageNet-1K [13] for our proposed self-supervised distillation, without using any annotations. After pre-training, we assessed the transfer learning capabilities of our method for downstream classification tasks on CIFAR-10 and CIFAR-100 [23], and for semantic segmentation on ADE20K [40]. We employed accuracy as the metric for image classification tasks and mean intersection over union (mIoU) as the metric for the semantic segmentation task. Detailed descriptions of the datasets can be found in the supplementary material.

Experiment Setup. We utilized the ViT-base [15] as the encoder for the student branch, with a patch size of 16 and 12 transformer blocks, each with 12-head multi-head attention and an embedding dimension of 768. The decoder in the student branch is a 2-layer shallow transformer with an embedding dimension of

Table 1. Comparison of top-1 fine-tuning accuracy on ImageNet-1K. CL denotes Contrastive Learning and MIM represents Masked Image Modeling

Method	Supervision	Acc (%)
Supervised [34]	Annotations	81.8
DINO [5]	CL	82.8
MoCo v3 [11]	CL	83.2
BEiT [3]	MIM	83.2
MAE [20]	MIM	83.6
SimMIM [37]	MIM	83.8
CIM [16]	MIM	83.3
CAE [9]	MIM	83.8
DMAE [2]	Combination	84.0
MimCo [41]	Combination	83.7
SiameseIM [33]	Combination	83.7
CAN [27]	Combination	83.6
MOMA	Combination	84.4

Table 2. Transfer learning top-1 accuracy (%) on CIFAR-10 and CIFAR-100.

Method	CIFAR-10	CIFAR-100
Random Init.	77.8	48.5
IN1K Sup. [15]	98.1	87.1
DINO [5]	99.1	91.7
MoCo v3 [11]	98.9	90.5
BEiT [3]	98.5	90.1
MAE [20]	99.1	91.6
SimMIM [37]	99.2	91.7
CAE [9]	99.1	91.7
DMAE [2]	99.2	91.6
SiameseIM [33]	99.1	91.6
CAN [27]	99.0	91.5
MOMA	99.2	91.8

512. The projector is a 2-layer MLP with a dimension of 768. We used the publicly available pre-trained ViT-base from MoCo v3 for the teacher branch, which shares the same architecture as the encoder in the student branch. During the self-supervised pre-training phase on ImageNet-1K, we trained the model for 800 epochs with a batch size of 1,024, using the AdamW [26] optimizer with a learning rate of $1.5e-4$, β_1 and β_2 set to 0.9 and 0.95, respectively, and a weight decay of 0.05. In the fine-tuning phase for classification tasks (including ImageNet-1K and transfer learning on CIFAR-10 and CIFAR-100), we fine-tuned the pre-trained ViT from the student encoder for 100 epochs with a batch size of 1024, using the AdamW optimizer with a learning rate of $1e-3$, β_1 and β_2 set to 0.9 and 0.999, respectively, and a weight decay of 0.05. For the semantic segmentation task, we integrated our pre-trained ViT into the UperNet [36] framework and fine-tuned it for 100 epochs with a batch size of 16, following the methodology outlined in [20]. All baseline comparison models adopted ViT-base as the backbone in the experimental results. Our experiments utilized 4 NVIDIA A100 GPUs during both the pre-training and fine-tuning phases.

4.2 Results on ImageNet-1K

We report the fine-tuning accuracy of our proposed MOMA framework alongside other baseline models on ImageNet-1K in Table 1. Our comparison includes models trained in a supervised manner, self-supervised approaches based on contrastive learning objectives, those utilizing masked image modeling objectives, and models employing a combination of different objectives. The results demonstrate that self-supervised baselines, including MOMA, surpass the supervised baseline, indicating that self-supervised pre-training on the large-scale ImageNet-1K dataset enhances fine-tuning classification performance. The rich and gener-

alized representations derived from self-supervised learning foster a more robust model that exceeds the performance of models trained solely on annotations. Furthermore, MOMA outperforms self-supervised baselines trained exclusively on either contrastive learning or masked image modeling objectives. This suggests that the synergistic integration of the contrastive learning teacher and the masked image modeling student yields superior representations for MOMA, leading to enhanced classification performance that leverages the strengths of both contrastive learning and masked image modeling. Compared to other self-supervised baselines that adopt a combination of different objectives, MOMA consistently outperforms them, achieving better performance than SiameseIM, Mimco, and CAN, which learn multiple objectives, including contrastive and masked reconstruction objectives. Additionally, MOMA surpasses DMAE, which utilizes knowledge distillation from a pre-trained large MAE. These findings indicate that MOMA effectively amalgamates knowledge from contrastive learning and masked image modeling. The knowledge from the off-the-shelf contrastive learning teacher reinforces global representations and high-level semantics in the masked image modeling student. Specifically, semantic-guided masks are more informative in capturing discriminative features and essential semantics than random masking. Moreover, the feature alignment process effectively enforces the global features within the masked image modeling student, leading to a more effective representation for vision tasks.

4.3 Transfer Learning on Downstream Tasks

Image Classification. We explored transfer learning for downstream classification tasks on CIFAR-10 and CIFAR-100. Both datasets are small-scale compared to the ImageNet-1K dataset, on which the proposed MOMA was self-supervised pre-trained. As illustrated in Table 2, we compared the proposed MOMA with a randomly initialized Vision Transformer (ViT) trained from scratch, a supervised ViT trained on ImageNet-1K, and various self-supervised baselines also pre-trained on ImageNet-1K. The randomly initialized ViT performed poorly on both datasets, likely due to the complex ViT model’s potential to overfit the small-scale data. In contrast, the supervised ViT trained on ImageNet-1K demonstrated excellent performance, showcasing that models trained on large-scale datasets can extract rich and powerful representations that transfer well to downstream tasks, even when the datasets are small-scale. Excluding DINO, all other self-supervised baselines surpassed both the supervised baseline and the randomly initialized baseline, highlighting the efficacy of self-supervised learned features in improving transfer learning ability and generalizability for downstream tasks. Among the self-supervised learning methods, MOMA consistently outperformed the other baselines on both datasets, indicating that MOMA acquires higher quality representations that generalize better on downstream classification tasks than other methods. The contrastive learning teacher within the proposed framework effectively enforces high-level semantics and global features onto the masked image modeling student, resulting in more generalizable learned representations with enhanced transfer learning capability.

Semantic Segmentation. We present the transfer learning semantic segmentation results in Table 3. The findings suggest that self-supervised learning generalizes well across different downstream vision tasks, achieving better performance than the supervised method. The proposed MOMA consistently outperforms both the supervised baseline and self-supervised baselines. The contrastive learning teacher and masked image modeling student within MOMA equip it with a combination of high-level semantic global features, as well as pixel-level representational power. Consequently, MOMA possesses enhanced discriminative power and has improved performance on pixel-level dense prediction tasks, achieving exceptional results in semantic segmentation.

Table 3. Transfer learning semantic segmentation results on ADE20K.

	Supervised [25]	DINO [5]	MoCo v3 [11]	BEiT [3]	MAE [20]	
mIoU (%)	46.6	47.2	47.3	48.8	48.1	
	SimMIM [37]	CAE [9]	DMAE [2]	SiameseIM [33]	CAN [27]	MOMA
mIoU (%)	50.0	50.1	49.7	49.6	48.8	50.2

4.4 Ablation Study

Choice of the Teacher Model. We considered various off-the-shelf pre-trained models as the teacher, including supervised models, masked image modeling pre-trained models, contrastive learning pre-trained models, and self-supervised pre-trained models that learned both contrastive and reconstruction objectives. As indicated in Table 4, the contrastive MoCo v3 pre-trained teacher model outperformed the supervised baseline and other self-supervised baselines. The contrastive learning pre-trained DINO model also demonstrated impressive results compared to other models. Supervised representations, coupled with supervision from classification annotations, and masked image modeling pre-trained models (MAE and SimMIM) that learn representations based on pixel-level reconstruction, do not incorporate critical features and global semantics. SiameseIM and CAN, despite being pre-trained with both contrastive and reconstruction objectives, do not effectively convey critical global features to the student model due to the neutralizing effect of the reconstruction objective. Thus, a pure contrastive teacher model is more effective in transferring knowledge to the student model, compensating for the lack of global features and high-level semantics in the masked image modeling pipeline.

Importance of Masking Strategy. We explored different masking strategies to investigate their importance in our proposed MOMA (see Fig. 2). Random masking [20] makes the masked image modeling task more challenging, potentially encouraging stronger models and better-learned representations. However,

these masks lack specific shapes or patterns and do not emphasize discriminative features or semantics. Block masking[3] creates rectangular masks with random sizes and aspect ratios, masking groups of neighboring pixels and being less scattered compared to random masking. Nonetheless, it does not take into account the semantic information during the mask generation process. In contrast, the proposed semantic-guided masking in MOMA leverages the attention map from the contrastive learning teacher, incorporating rich semantics, including discriminative features of the input data. According to the results in Table 5, semantic-guided masking outperforms the other two strategies, indicating that guidance from the contrastive learning teacher encourages the masked image modeling to focus more on discriminative features and semantics, leading to superior representations and better performance in vision tasks.

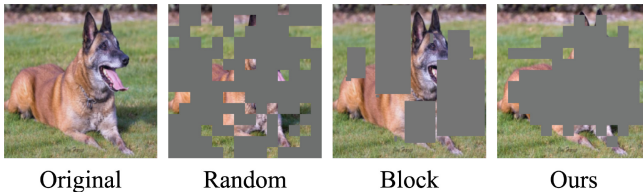


Fig. 2. Comparison of Different Masking Strategies. We illustrate the comparison of different masking strategies. From left to right, the figures display the original input image, followed by random masking [20], block masking [3], and our proposed semantic-guided masking.

Impact of Feature Alignment. The goal of feature alignment is to bring the student representations from the masked image modeling pipeline closer to the target representations produced by the contrastive learning teacher. This alignment allows the student’s representations to better match the teacher’s, which are rich in high-level semantics and global features. As shown in Table 6, removing feature alignment significantly decreased MOMA’s performance, underscoring the importance of global features and semantics from the teacher model in generating beneficial representations for learning. Additionally, we explored different alignment functions within the framework, including cosine similarity, L_1 , and L_2 metrics. All alignment functions improved performance compared to the absence of feature alignment, highlighting the critical role of feature alignment in capturing global semantics, and improving representation quality. The smooth L_1 metric achieved better performance than other functions, as it is more robust, stable, and easier to optimize during learning [32].

Table 4. Comparison of top-1 fine-tuning accuracy on ImageNet-1K regarding choice of teacher.

Teacher	Acc (%)
Supervised [34]	83.7
DINO [5]	84.1
MoCo v3 [11]	84.4
MAE [20]	83.8
SimMIM [37]	83.9
SiameseIM [33]	84.0
CAN [27]	83.8

Table 5. Comparison of top-1 fine-tuning accuracy on ImageNet-1K regarding masking strategies.

Masking	Acc (%)
Random [20]	84.1
Block [3]	83.8
Semantic-guided	84.4

Table 6. Comparison of top-1 fine-tuning accuracy on ImageNet-1K regarding feature alignment methods.

Alignment	Acc (%)
None	83.7
Cosine Similarity	84.2
L_1 Distance	84.3
L_2 Distance	84.2
Smooth L_1	84.4

5 Conclusion

In this work, we introduced MOMA, a self-supervised knowledge distillation framework that enhances masked image modeling by encouraging the integration of high-level semantics and global features. We leveraged a contrastive learning teacher to generate semantic-guided masks for the masked image modeling student and incorporated feature alignment between the teacher and student representations. Our extensive experiments across various vision benchmarks demonstrate that the proposed MOMA framework achieves improved performance compared to traditional masked image modeling frameworks and exhibits competitive results relative to other self-supervised baselines. We aim to apply the proposed method to other critical domains, such as medical applications and aspire to inspire advancements in the design of more sophisticated self-supervised learning algorithms.

Acknowledgements. This research was supported by The University of Melbourne’s Research Computing Services and the Petascale Campus Initiative.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint [arXiv:1607.06450](https://arxiv.org/abs/1607.06450) (2016)
2. Bai, Y., Wang, Z., Xiao, J., Wei, C., Wang, H., Yuille, A., Zhou, Y., Xie, C.: Masked autoencoders enable efficient knowledge distillers. arXiv preprint [arXiv:2208.12256](https://arxiv.org/abs/2208.12256) (2022)
3. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint [arXiv:2106.08254](https://arxiv.org/abs/2106.08254) (2021)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)

5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
6. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International conference on machine learning. pp. 1691–1703. PMLR (2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. *Adv. Neural. Inf. Process. Syst.* **33**, 22243–22255 (2020)
9. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning. arXiv preprint [arXiv:2202.03026](https://arxiv.org/abs/2202.03026) (2022)
10. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning. *Int. J. Comput. Vision* **132**(1), 208–223 (2024)
11. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
12. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv preprint [arXiv:1602.02830](https://arxiv.org/abs/1602.02830) (2016)
13. 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)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
16. Fang, Y., Dong, L., Bao, H., Wang, X., Wei, F.: Corrupted image modeling for self-supervised visual pre-training. arXiv preprint [arXiv:2202.03382](https://arxiv.org/abs/2202.03382) (2022)
17. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: Seed: Self-supervised distillation for visual representation. arXiv preprint [arXiv:2101.04731](https://arxiv.org/abs/2101.04731) (2021)
18. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
19. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural. Inf. Process. Syst.* **33**, 21271–21284 (2020)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)

22. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint [arXiv:1503.02531](https://arxiv.org/abs/1503.02531) **2**(7) (2015)
23. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint [arXiv:1711.05101](https://arxiv.org/abs/1711.05101) (2017)
27. Mishra, S., Robinson, J., Chang, H., Jacobs, D., Sarna, A., Maschinot, A., Krishnan, D.: A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. arXiv preprint [arXiv:2210.16870](https://arxiv.org/abs/2210.16870) (2022)
28. Noroozi, M., Vinjimoor, A., Favaro, P., Pirsivash, H.: Boosting self-supervised learning via knowledge transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9359–9367 (2018)
29. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint [arXiv:1807.03748](https://arxiv.org/abs/1807.03748) (2018)
30. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
31. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint [arXiv:2010.04592](https://arxiv.org/abs/2010.04592) (2020)
32. Sutanto, A.R., Kang, D.K.: A novel diminish smooth l1 loss model with generative adversarial network. In: Intelligent Human Computer Interaction: 12th International Conference, IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part I 12. pp. 361–368. Springer (2021)
33. Tao, C., Zhu, X., Su, W., Huang, G., Li, B., Zhou, J., Qiao, Y., Wang, X., Dai, J.: Siamese image modeling for self-supervised vision representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2132–2141 (2023)
34. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
35. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
36. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
37. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
38. Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. *Advances in neural information processing systems* **32** (2019)
39. Yao, Y., Desai, N., Palaniswami, M.: Masked contrastive representation learning. arXiv preprint [arXiv:2211.06012](https://arxiv.org/abs/2211.06012) (2022)

40. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralla, A.: Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vision* **127**(3), 302–321 (2019)
41. Zhou, Q., Yu, C., Luo, H., Wang, Z., Li, H.: Mimco: Masked image modeling pre-training with contrastive teacher. In: *Proceedings of the 30th ACM International Conference on Multimedia*. pp. 4487–4495 (2022)



Leveraging Cross-Augmentation Consensus and Conflict for Semi-supervised Semantic Segmentation

Junhao Cao¹, Junyi Chen¹, Sib0 Huang², and Dongyu Zhang¹(✉)

¹ School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

zhangdy27@mail.sysu.edu.cn

² Network and Information Center, Huizhou University, Huizhou, China

Abstract. Semi-supervised semantic segmentation leverages both labeled and unlabeled images to accomplish pixel-wise classification task. Within this field, the weak-to-strong consistency regularization has been widely popularized and has become a standard approach. However, unidirectional regularization often leads to the ignorance of correct but filtered predictions and brings the noise of wrong but confident predictions. To address these inherent flaws, we fully leverage Cross-Augmentation Consensus and Conflict (CACC), including Augmentation Feedback Mechanism (AFM) and Category Threshold Controller (CTC). AFM aims to mitigate the influence of incorrect predictions with high-confidence and mine unconfident but accurate predictions by re-weighting the pixel-wise pseudo supervision and applying supplementary regularization. Concurrently, CTC adopts category-specific thresholds by considering the model's overall performance and the varying category-specific learning difficulty. Experimental results on benchmark datasets demonstrate the superior performance of our method, showcasing its effectiveness in improving semi-supervised semantic segmentation.

Keywords: Semi-Supervised Learning · Semantic Segmentation

1 Introduction

Semantic segmentation is a fundamental task in computer vision, which is about classifying each pixel in an image into semantic categories. It is widely applied in various visual fields, including autonomous driving [12], medical image analysis [23] and remote-sensing image analysis [16]. Although it is crucial, densely per-pixel labeling is time-consuming and labor-intensive. Considering the substantial

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78398-2_6.

demand for annotations, semi-supervised semantic segmentation emerges as a practical solution. This approach leverages a limited set of labeled data alongside a larger pool of unlabeled images, which not only alleviates the annotation burden but also enhances the generalization of the segmentation model.

Specifically, semi-supervised semantic segmentation can be classified into two principal approaches. The first one is the pseudo-labeling, which assigns high-quality pseudo-labels to unlabeled data, thereby transferring the knowledge from labeled to unlabeled data [11]. The second one is the consistency regularization, which enforces the model to produce stable outputs for perturbed inputs [18].

A foundational work in this area is FixMatch [20]. Specifically, it generates pseudo-labels from the weakly augmented images and uses these pseudo-labels to supervise the predictions of strongly augmented images, which also serves as the consistency regularization among different perturbed views. Unlike many other semi-supervised learning methods, FixMatch does not rely on complex auxiliary components or post-processing workflows. FixMatch has inspired subsequent innovations including UniMatch [29], CorrMatch [21] and MaskMatch [19].

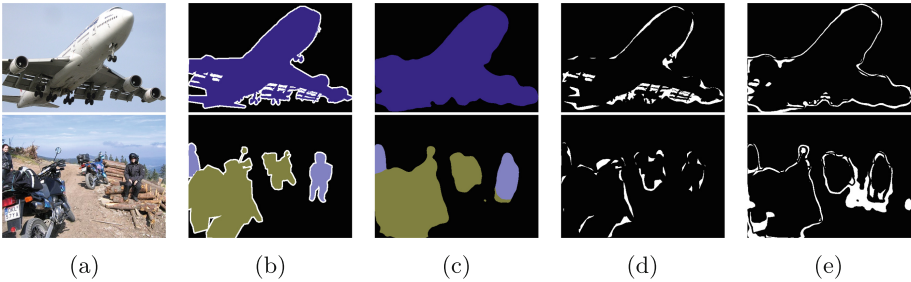


Fig. 1. Qualitative results generated by UniMatch [29] on the Pascal VOC 2012 dataset. We fix 0.9 as the threshold. (a) Images; (b) Ground truth; (c) Predictions; (d) High-confidence wrong predictions (marked as white); (e) Low-confidence correct predictions (marked as white).

However, despite their simplicity and efficiency, FixMatch-family methods are not without its limitations. As column (d) in Fig. 1 shows, predictions with high-confidence from weakly perturbed images as pseudo-labels may introduce noise into the training process, as some of these reliable predictions may be erroneous. This noise can accumulate over the course of training, leading to the confirmation bias of the model. As column (e) in Fig. 1 shows, a non-negligible fraction of pixels are discarded due to threshold-based filtering, potentially omitting valuable information that could contribute to the model learning.

Additionally, the one-size-fits-all threshold overlooks the overall training performance of the model and class-wise learning difficulties. High thresholds may filter out potentially correct predictions for challenging categories and inhibits their learning. Low thresholds may introduce noise from easier-to-learn categories predictions, exacerbating the model’s bias towards these categories.

Based on the analysis above, this paper attempts to address the following issues without introducing additional components, **solely based on the conflicts and consensus among different augmentation flows within the model itself**: (i) *How to reduce the impact of reliable but incorrect predictions?* (ii) *How to re-mine unreliable but correct predictions, treating filtered “trash” as “treasure”?* (iii) *How to design dynamically adjusted category thresholds that consider the global and class-wise learning progress?* To this end, the paper proposes the Augmentation Feedback Mechanism (AFM) to tackle the issues of model overreaction (problem i) and underreaction (problem ii), while the Category Threshold Controller (CTC) addresses the issue of category learning imbalance caused by fixed model thresholds (problem iii). **For AFM**, robust supervision from the consensus among data augmentations is selected as a supplementary supervision when predictions from the weak augmentation branch are missing due to the threshold filtering. Adaptive weight adjustment is also applied to the weak augmentation branch predictions based on consensus within the augmentation space. **For CTC**, the strategy is based on the following hypothesis: categories that are close in the representation space are easily confused after applying data augmentation [14]. Therefore, CTC calculates the transition frequency among categories between weak and strong augmentation, which is a measure of confusion among categories. Category thresholds are determined based on the degree of confusion across categories, thus setting flexible constraints for hard-to-learn samples and strict constraints for easy-to-learn samples.

Overall, the contributions of this work are summarized into three folds:

- This paper proposes AFM to reduce the impact of reliable but incorrect predictions and to utilize unreliable but correct predictions.
- This paper proposes CTC, which dynamically adjusts class-wise thresholds based on the conflict among categories.
- Experiments on extensive benchmarks has demonstrated the superior performance of the proposed method.

2 Related Work

As outlined in Sec. 1, semi-supervised learning bridges the gap between supervised and unsupervised methodologies by utilizing both labeled and unlabeled data. This approach leads to two key strategies: consistency regularization and pseudo-label training. Consistency regularization focuses on producing stable outputs under variations of image space or feature space. In contrast, pseudo-label training assigns proxy labels to unlabeled data, thereby continuously guiding the model towards a more supervised pattern.

2.1 Consistency Regularization

Consistency regularization capture the distributional structure of unlabeled data, thereby enhancing the model’s generalization. Some works focus on apply-

ing adaptive strong augmentations to enhance data diversity. [33] enhances semi-supervised semantic segmentation performance through intensity-based augmentations and adaptive CutMix [31] techniques. [32] introduces instance-specific enhancements and model-adaptive supervision to address instance learning-difficulties variability. [9] tackles class imbalance by employing adaptive Copy-Paste and CutMix augmentations and a re-weighting strategy to balance category performance. Some works explore a broad perturbation space. [29] explores image and feature perturbation spaces and integrates them into a unified framework. [15] presents a dual-teacher framework which jointly injects feature-level adaptive perturbations to the student model.

In this work we explore a noise-resistant consistency regularization strategy through the interactions between data augmentation flows within the model itself, without relying on additional components or auxiliary contrastive loss.

2.2 Pseudo labeling

Pseudo-label self-training methods use pseudo-labels to convert unlabeled data into the annotated format, reducing the gap between semi-supervised and fully-supervised approaches. For rectification, [4] proposes a decoupling training strategy and an entropy-based sampling strategy to train a class-unbiased decoder. [17] adopts dynamic soft pseudo labels to maintain the potential ground-truth classes. For cross-model supervision, [13] applies a cross-fusion supervision mechanism to fuse predictions from multiple learners, as well as applying a lower weight to the object boundary to mitigate the noise from unreliable pixels. [6] trains two parallel classifiers by the supervision of the intersection and union between their predictions, and [26] minimizes the similarity between the feature extracted by two sub-nets. Both of them encourage to learn reliable predictions from two irrelevant views for co-training.

In this work, we do not rely on a multi-model system to generate complementary and robust pseudo-labels. Instead, we construct dynamically adjusted thresholds based on the conflicts between data augmentation flows within the model. This approach can filter out unreliable pseudo-labels while retaining high-quality ones.

3 Methodology

In this section we delve into the proposed CACC approach with AFM and CTC. Fig. 2 shows the overall pipeline of our method. In Sec. 3.1, we briefly define the task and introduce the motivation of the proposed CACC. In Sec. 3.2 we introduce AFM to dynamically adjust prediction weights and incorporate additional supervision signals based on augmentation consensus. In Sec. 3.3 we introduce CTC, focusing on how to adjust category thresholds based on cross-category conflicts.

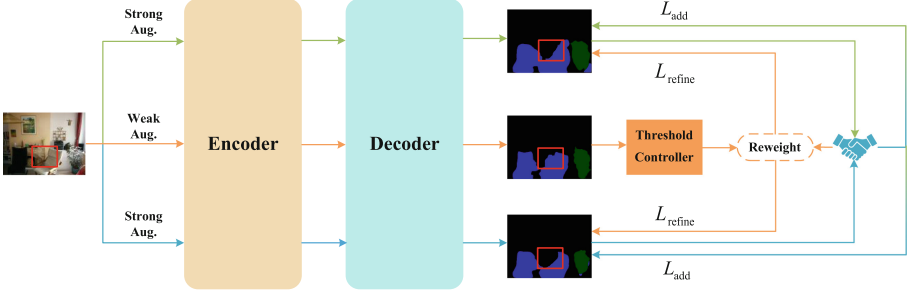


Fig. 2. The overview pipeline of the proposed CACC. Images with weak augmentation (arrows marked as orange) are sent to the model and the pseudo labels are generated to supervise predictions from two strong augmentation flows (arrows marked as green and blue). The original weak-to-strong loss will be modified into L_{refine} by reweighting strategy (Sec. 3.2) and threshold controller (Sec. 3.3). Beyond the $L_{weak-strong}$, an additional loss L_{add} will serve as the supplementary supervision based on the prediction consensus (Sec. 3.2).

3.1 Preliminary

Semi-supervised learning combines supervised and unsupervised learning methods to improve model performance. In supervised learning, the model is trained on a labeled dataset to learn the prior classification information. In unsupervised learning, the model uses an unlabeled dataset to identify patterns of the data. Using a larger amount of unlabeled data besides the labeled one often results in better generalization and accuracy.

To be more specific, given a dataset $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$, where $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{N_L}$ represents the labeled set with N_L image-label pairs and $\mathcal{D}_U = \{x_j\}_{j=1}^{N_U}$ represents the unlabeled set with N_U images, the goal is to classify each pixel into K categories with a limited amount of labeled data alongside a larger set of unlabeled data. The overall loss function can be described as Eq. 1

$$\mathcal{L} = \mathcal{L}_{\text{unsup}} + \lambda \mathcal{L}_{\text{sup}}, \quad (1)$$

where λ is a trade-off factor to balance two targets and $\mathcal{L}_{\text{sup}} = -\frac{1}{N_L} \sum_{j=1}^{N_L} H(\mathbf{y}_j, \mathbf{p}_j)$.

For general setting, pseudo-labels generated from weakly augmented version is used to guide the learning on strongly augmented data. The unsupervised loss function can be written as Eq. 2

$$\mathcal{L}_{\text{unsup}} = -\frac{1}{N_U} \sum_{j=1}^{N_U} \mathbf{M}_j \odot H(\hat{\mathbf{y}}_j^w, \mathbf{p}_j^s), \quad (2)$$

where N_U is the number of unlabeled images. $\mathbf{M}_j = \mathbb{I}(\mathbf{p}_j^w > \tau)$ is the indicator matrix for the j -th image, where \mathbb{I} is the indicator function, returning 1 when the condition $\max(\mathbf{p}_j^{\text{weak}}) > \tau$ is satisfied (*i.e.*, when the prediction probability

exceeds the confidence threshold τ) and 0 otherwise for each pixel. \odot represents the Hadamard product. H is the cross-entropy function. $\hat{\mathbf{y}}_j^w$ is the one-hot label for the weakly augmented view, and \mathbf{p}_j^s is the prediction probability matrix for the strongly augmented versions.

Upon revisiting this equation, we encounter three issues: First, predictions from the weakly augmented branch are not always reliable, and it is needed to identify these incorrect predictions. Second, it is important to retrieve the accurate predictions filtered by the threshold. Third, the threshold should be dynamically adjusted to accurately distinguish erroneous and correct samples as the training goes on. Sec. 3.2 will delve into the first two issues, while the discussion on the third issue will be presented in Sec. 3.3.

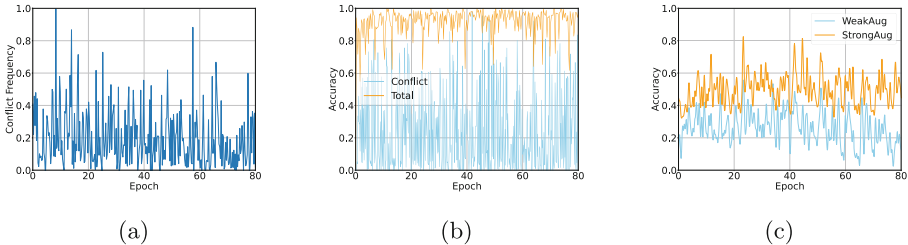


Fig. 3. Supportive experiments of CACC. Fig. 3(a) shows that the phenomenon exists where predictions of weak and strong augmented images don’t reach a consensus. Fig. 3(b) shows that the accuracy of pseudo labels in conflict is lower than the normal one. Fig. 3(c) shows that the consensus predictions from both strong-augmented flows are more accurate where the pseudo labels are filtered.

3.2 Augmentation Feedback Mechanism

This chapter primarily investigates how to enhance the utilization of pseudo-labels in unlabeled datasets and the capability to filter noisy pseudo-labels. As mentioned in Sec. 2.2, cross supervision helps mitigate confirmation bias and promote knowledge exchange. In fact, the concept of cross-supervision among multiple models can also be transferred to a single model by considering different data augmentation branches as views of different models. Consequently, a straightforward idea emerges that the consensus among predictions from multiple strongly augmented branches is reliable. This consensus can not only serve as the judgement to the accuracy of predictions from weakly augmented branches, but also be a supplementary supervisory signal when the pseudo-supervision from weakly augmented branches is filtered by the threshold.

To verify the reliability of this idea, we conduct experiments on the Pascal VOC *blender* [7] 1/8 split with ResNet-101 [8] backbone. First, we select all pixels where predictions from dual augmentation branches are consistent, and then we extract predictions of selected pixels from the weak augmentation branch.

We measure the frequency of conflicts between the prediction from the weak and strong augmentation branch, and calculate the proportion of pixels whose predictions from the weak augmentation branch is wrong but the predictions from the strongly augmented one is correct. Fig. 3(a) indicates that pixels with conflict occurs in high frequency and needs consideration. Fig. 3(b) indicates that pixels with conflicts are more likely to exhibit noisy predictions compared to regular pixels. Second, we fix thresholds $\tau = 0.8$ to identify pixels filtered out by the thresholds. We calculate the accuracy of these filtered pixels where the predictions from the strong augmentation branches agree. The results in Fig. 3(c) demonstrate that these consistent predictions could indeed serve as a form of reliable auxiliary supervision.

The experimental results strongly support our hypothesis that the consensus among multiple augmented views and the discrepancies in weak-strong view pairs serve as a robust indicator for identifying errors in pseudo-labels. Based on this, Eq. 2 can be modified. Specifically, we update the mask as Eq. 3:

$$\mathbf{N}_j = \begin{cases} 1, & \text{if } \hat{\mathbf{y}}_j^{s1} \neq \hat{\mathbf{y}}_j^{s2} \\ \max(\mathbf{p}_j^w), & \text{if } \hat{\mathbf{y}}_j^{s1} = \hat{\mathbf{y}}_j^{s2}. \end{cases} \quad (3)$$

For the part where $\mathbf{N}_j = 1$, the strong augmentation branches fail to reach a consensus, and their predictions are probably unreliable. It is essential to rely on the supervision provided by the weak augmentation branch. For the part where $\mathbf{N}_j < 1$, pixels can be divided into two subsets: First, in cases where conflicts arise between weak and strong augmented predictions, dynamic weighting is applied to the loss function. A lower confidence in the weakly augmented branch indicates that the pseudo-prediction is less reliable, thereby mitigating the noise from pseudo-labels. Second, if the predictions from the weak and strong branches are consistent, it suggests that the model has adequately learned the information for the region. As such, further learning is unnecessary, and the model’s attention on these reliable regions can be reduced by diminishing their weights.

Beyond the vanilla loss function, we add a supplement term where predictions from weakly augmented view are filtered. Similar to Eq. 3, we define the weight as Eq. 4:

$$\mathbf{R}_j = \begin{cases} 0, & \text{if } \hat{\mathbf{y}}_j^{s1} \neq \hat{\mathbf{y}}_j^{s2} \\ 1 - \max(\mathbf{p}_j^w), & \text{if } \hat{\mathbf{y}}_j^{s1} = \hat{\mathbf{y}}_j^{s2}. \end{cases} \quad (4)$$

Overall, mitigating the concept of Eq. 3 and Eq. 4, Eq. 2 can be re-formulated to Eq. 5:

$$\begin{aligned} \mathcal{L}_{\text{refine}} &= -\frac{1}{2N_U} \sum_{j=1}^{N_U} \mathbf{M}_j \odot \mathbf{N}_j \odot (H(\hat{\mathbf{y}}_j^w, \mathbf{p}_j^{s1}) + H(\hat{\mathbf{y}}_j^w, \mathbf{p}_j^{s2})), \\ \mathcal{L}_{\text{add}} &= -\frac{1}{2N_U} \sum_{j=1}^{N_U} (1 - \mathbf{M}_j) \odot \mathbf{R}_j \odot (H(\hat{\mathbf{y}}_j^{s2}, \mathbf{p}_j^{s1}) + H(\hat{\mathbf{y}}_j^{s1}, \mathbf{p}_j^{s2})), \\ \mathcal{L}_{\text{unsup}} &= \mathcal{L}_{\text{refine}} + \mathcal{L}_{\text{add}}. \end{aligned} \quad (5)$$

3.3 Category Threshold Controller

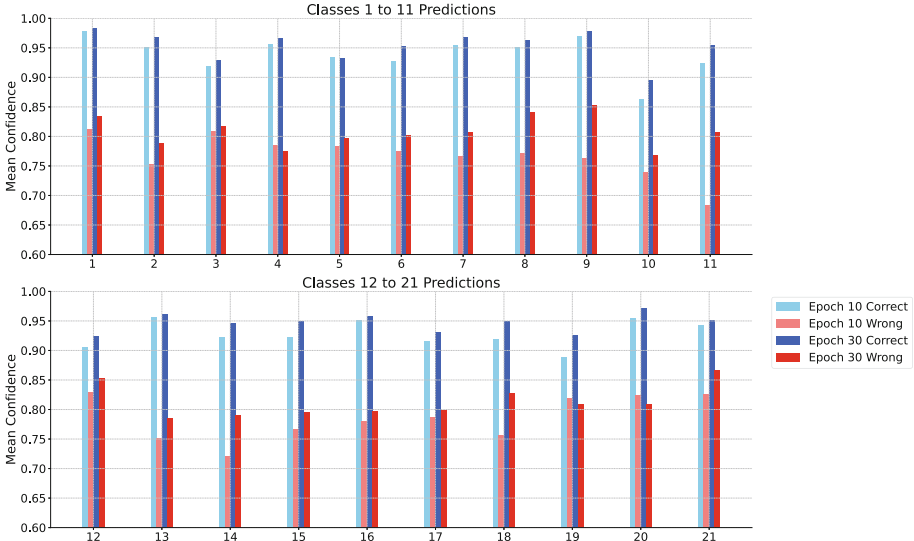


Fig. 4. Comparison of mean confidence of correct predictions and wrong predictions during 10th and 30th Epoch across Pascal VOC 2012 classes (with ResNet-101 model).

Before designing our threshold strategy, we firstly compute the mean value of confidence and the corresponding standard deviation¹ of correct and incorrect predictions for each category during the 10th and 30th epochs (corresponds to the captions of "correct" and "wrong" in Fig. 4). Two key insights can be obtained from Fig. 4: First, the overall performance is improving, indicating a gradual enhancement in model performance. Second, the confidence level vary among categories. There exist easy-to-learn and difficult-to-learn categories. Such observation have inspired our approach to designing thresholds: (1) they should reflect the overall learning progress of the model, and (2) the thresholds should be category-specific. Nearly all previous studies employ a fixed threshold, which is contradictory to the first principle. A few works including CorrMatch [21] and U²PL [25], despite obeying to principle 1, overlook the principle 2.

Building upon the above discussions, we compute both the global threshold and category confusion factors to reflect the model’s overall training performance and the learning difficulty of categories. For the global threshold, we calculate the online global threshold for the current batch by averaging the confidence of all pixels exceeding the current global threshold, and then update the historical global threshold using exponential moving average (EMA):

$$\bar{P} = \frac{\sum_i \max(p_i^w) \cdot \mathbb{I}(\max(p_i^w) > \tau_g^{t-1})}{\sum_i \mathbb{I}(\max(p_i^w) > \tau_g^{t-1})}, \quad (6)$$

¹ The standard deviation will be provided in Sec. A in the supplementary material.

$$\tau_g^t = \alpha \cdot \tau_g^{t-1} + (1 - \alpha) \cdot \bar{P}, \quad (7)$$

where τ_g means the global threshold, t means the time step, \bar{P} means the online factor, i indicates the pixel, α is the momentum decay of EMA. Secondly, we compute the confusion factor for each category to represent the learning difficulty of each category. Specifically, We initialize a confusion matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$ and count the transfer frequency in a minibatch. $\mathbf{C}[r, c]$ means the number of pixels where the weak augmentation predictions belong to category r and either of the strong augmentation predictions belong to category c . Then we sum the rows of matrix \mathbf{C} and divide each diagonal element by the row-wise sum to obtain the normalized category confusion factors $\bar{F}_k^t \in \mathbb{R}^K$:

$$\bar{F}_k^t = \frac{\mathbf{C}_{kk}}{\sum_c \mathbf{C}_{kc}}, \quad (8)$$

where k represents the index of category. Like the global threshold, we also update the category confusion factors using EMA:

$$F_k^t = \alpha \cdot F_k^{t-1} + (1 - \alpha) \cdot \bar{F}_k^t. \quad (9)$$

Finally, we normalize the category confusion factors to their maximum value to obtain the relative confusion factors for each category, which are then multiplied by the global threshold to determine the final category-specific thresholds:

$$\tau_k^t = \tau_g^t \cdot \frac{F_k^t}{\max(F_k^t)}. \quad (10)$$

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on both the PASCAL VOC 2012 [5] and Cityscapes [2] datasets to evaluate our method. The PASCAL VOC 2012 dataset comprises 1,464 finely annotated training images and 1,449 validation images. Additionally, we incorporate additional 9,118 coarsely labeled images from the Semantic Boundaries Dataset (SBD) [7], the same as all the baseline methods mentioned in Tab. 2. We evaluate our method on both *classic* and *blender* Pascal setting. For the Cityscapes dataset, it is an urban-scene-related dataset, including 2,975 images for training and 500 images for validation. For all datasets we adopt 1/2, 1/4, 1/8, and 1/16 labeled data ratio as the settings of experiments.

Evaluation Protocols. For all datasets, we use the mean Intersection-over-Union (mIoU) as our evaluation metric. During the inference phase, for the Pascal dataset, we center-crop the image to a fixed size and conduct inference for whole cropped image. For the Cityscapes dataset, we adopt the sliding-window way for image inference. Inferences are conducted on the validation sets of all datasets.

Implementation Details. For the batch size, there are 8 labeled and 8 unlabeled images in a minibatch. For the learning rate, we start with 0.001 for the Pascal dataset and 0.005 for the Cityscapes dataset. The learning rate for the decoder is set 10 times as large as that of the backbone network for Pascal. The optimization is carried out by an SGD optimizer with a weight decay of 0.0001, and we employ a poly policy to adjust learning rate. For the image size, we resize the Pascal dataset images to 512×512 and Cityscapes images to 769×769 . All augmentation methods keep the same as UniMatch and all derived works. We set 80 epochs and 240 epochs of the training process for Pascal and Cityscapes. For the model we use DeepLab v3+ [1] with the output stride of 16 as the segmentation network, and the backbone is ResNet101 [8] pretrained on ImageNet [3]. All experiments are conducted on 4 V100 GPUs.

4.2 Comparison with State-of-the-Arts methods

Table 1. Comparison with the state-of-the-art methods on Pascal VOC 2012 *Classic* Val set. † means that since the released code corresponds to the older version of the manuscript on arXiv, we only report the previous results. See the paper for more details. The highest mIOU is marked in red, and the second highest mIOU is marked in blue. Same as below.

Methods	Venue	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)
SupOnly	-	44.98	50.79	63.88	69.30
PRCL [27]	AAAI23	69.91	74.42	76.69	77.88
MKD† [30]	ACMMM23	65.35	70.18	74.44	75.90
AugSeg [33]	CVPR23	71.09	75.45	78.80	80.33
CCVC [26]	CVPR23	70.20	74.40	77.40	79.10
DGCL [24]	CVPR23	70.47	77.14	78.73	79.23
iMAS [32]	CVPR23	68.80	74.40	78.50	79.50
UniMatch [29]	CVPR23	75.20	77.20	78.80	79.90
CSS [22]	ICCV23	68.09	71.93	74.91	77.57
ESL [17]	ICCV23	70.97	74.06	78.14	79.53
DeS ⁴ [4]	IJCAI23	68.02	72.23	74.58	77.62
PCR [28]	NIPS22	70.06	74.71	77.16	78.49
GTA [10]	NIPS22	70.02	73.16	75.57	78.37
Ours	-	75.44	77.75	79.28	80.48

Results on classic Pascal VOC 2012 dataset. As shown in Tab. 1, our method significantly outperforms the SupOnly baseline by 30.46%, 26.96%, 15.40% and 11.18%, showing an impressive improvement across all subsets, with the most notable increase being over 30% in the 1/16 subset. When compared to

Table 2. Comparison with the state-of-the-art methods on Pascal VOC 2012 *Blender* Val set. [†] is reproduced with the output stride of 16. See the [code](#) for more details. 1/2 result of [§] is reproduced by abandoning the Dropout in the backbone and keeping the original dilation rate, as same as all baseline methods and ours. Since U²PL prioritizes selecting high quality labels in *blender* experiment setting, we compare with methods using the same split as U²PL for fairness (marked as[†]).

Methods	Venue	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupOnly	-	67.26	69.05	75.03	76.81
CorrMatch [†] [21]	CVPR24	77.82	78.57	78.96	-
MKD [30]	ACMMM23	75.90	76.59	77.62	78.94
MaskMatch [19]	Arxiv23	76.66	78.56	79.44	-
AugSeg [33]	CVPR23	77.01	77.31	78.82	-
CCVC [26]	CVPR23	77.20	78.40	79.00	-
DGCL [§] [24]	CVPR23	76.61	78.37	79.31	79.87
iMAS [32]	CVPR23	76.50	77.90	78.10	-
UniMatch [29]	CVPR23	78.10	78.40	79.20	-
ESL [17]	ICCV23	76.36	78.57	79.02	79.98
Ours	-	78.50	78.98	79.50	80.03
U ² PL [†] [25]	CVPR22	77.20	79.00	79.30	-
CSS [†] [22]	ICCV23	78.73	79.54	80.82	81.06
GTA [†] [10]	NIPS22	77.82	80.47	80.57	81.01
PCR [†] [28]	NIPS22	78.60	80.71	80.78	80.91
Ours [†]	-	80.81	81.64	81.77	81.97

existing SOTA methods, our approach surpasses leading methods such as UniMatch and AugSeg. Specifically, in the constrained subset (1/8 labeled ratio), our method exceeds the SOTA method by 0.55%, demonstrating our method’s effectiveness in leveraging limited labeled data.

Results on *blender* Pascal VOC 2012 dataset. As shown in Tab. 2, our method again demonstrates superior performance over the SupOnly baseline, with up to a 13.54% increase in mIOU in the 1/16 subset. Against the current SOTAs, our method outperforms them by 0.40%, 0.41%, 0.06% and 0.05%, respectively. Under the setting of U²PL, our method exceeds the baseline methods by 2.08%, 0.93%, 0.95% and 0.91% respectively. Two groups of experiments demonstrate that our method perform well under different partitions.

Results on Cityscapes dataset. In this dataset, known for its complexity due to the diversity of urban scenes, our method continues to exhibit improvements over the SupOnly baseline, especially with a 9.82% increase in the 1/16 subset. Compared to leading SOTAs, our method maintains a competitive outperforming by 0.17%, 0.23%, 0.11% respectively. Under 1/2 split our method still achieves comparable result.

Table 3. Comparison with the state-of-the-art methods on Cityscapes Val set. [†]is reproduced with the output stride of 16 and the resolution of 769, see the [code](#) for more details. [‡]are reproduced with the resolution of 769 by us and MaskMatch, respectively.

Methods	Venue	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
SupOnly	-	66.30	72.80	75.00	78.00
CorrMatch [†] [21]	CVPR24	75.95	77.63	78.27	79.34
MKD [30]	ACMMM23	75.31	75.98	78.28	80.74
MaskMatch [19]	Arxiv23	75.68	77.82	78.71	80.29
AugSeg [‡] [33]	CVPR23	74.93	77.42	78.77	79.61
CCVC [26]	CVPR23	74.90	76.40	77.30	-
DGCL [24]	CVPR23	73.18	77.29	78.48	80.71
iMAS [32]	CVPR23	74.30	77.40	78.10	79.30
UniMatch [‡] [29]	CVPR23	75.76	77.61	78.60	79.08
CSS [22]	ICCV23	74.02	76.93	77.94	79.62
ESL [17]	ICCV23	75.12	77.15	78.93	80.46
DeS ⁴ [4]	IJCAI23	-	75.74	77.87	-
PCR [28]	NIPS22	73.41	76.31	78.40	79.11
Ours	-	76.12	78.05	79.04	79.82

4.3 Ablation Study

Table 4. Ablation study of the proposed components in CACC.

AFM _{add}	AFM _{refine}	CTC	Pascal(183)	Pascal(1/8)	Cityscapes(1/2)
			75.90	77.57	78.73
	✓		76.78	78.23	79.64
✓	✓		77.19	78.44	79.82
		✓	76.57	78.16	78.73
✓	✓	✓	77.75	78.98	79.82

We conduct the ablation study on the *classic* Pascal 183 split, *blender* Pascal 1/8 split and Cityscapes 1/2 split to provide insightful observations on the individual and combined effects of AFM and CTC within the CACC framework.

Effectiveness of AFM. Initially, the baseline model (with only two branches of strong augmentations) achieves performance scores of 75.90%, 77.57%, and 78.73% mIOU respectively. The inclusion of the adaptive weighting component (AFM_{refine}) demonstrates independently beneficial effects, with increases of 0.88%, 0.66% and 0.91%. This demonstrates its efficacy in handling diverse

Table 5. Comparison of Computational Burden under the setting of the classic Pascal 732 split.

Method	GFLOPs	Time (per epoch)
Baseline	2197.41	19min37s
CCAC	2197.41	19min39s

data scenarios. The simultaneous application of both AFM components further enhanced model performance to 77.19%, 78.44%, and 79.82%. This result indicates the complementary effect of two components. Notice that just adopting AFM_{add} individually will lead to the failure of training, since the majority of loss function is still $\text{AFM}_{\text{refine}}$.

Effectiveness of CTC. Notice that since we adopt zero as the threshold, following the setting of all baseline methods, there is no contribution of CTC in Cityscapes. Compared to the fixed threshold, the integration of CTC further yields the improvement of 0.56% and 0.54%, which verifies its pivotal role in dynamically balancing learning thresholds based on category confusion, thereby distinguishing the correct and wrong predictions. For the contribution to the class-wise IOU, please refer to the Sec. B in the supplementary material.

Ablation Study on Hyper-parameters. The only hyper-parameter in the proposed method is the momentum factor α . Please refer to Sec. C in the supplementary material.

Analysis on the computational burden. Although incorporating two components in our method, it can be concluded that the additional computational burden is negligible. From the perspective of the theoretical aspect, the additional computational burden only comes from a few matrices. In Eq. 5, the additional matrices are \mathbf{R}_j and \mathbf{N}_j . The extra computation includes the argmax-operation and comparison between $\hat{\mathbf{y}}_j^{\text{s1}}$ and $\hat{\mathbf{y}}_j^{\text{s2}}$, and the calculation of per-pixel weight. The amount of computation is $N_u \times (2 \times K + 1) \times H \times W$, where H and W are the width and height of an image. In Eq. 10, the extra computation comes mainly from the matrix operations, including Eq. 6 and the construction of matrix \mathbf{C} . And the computation can be estimated as $2 \times N_u \times H \times W$. Overall, just a few image-level matrices computation is totally negligible, compared to the large amount of matrices computation in the neural network. We also conduct the experiment to compare the computational burden between our baseline method and the CACC method. As shown in Tab. 5, adding our designed components will not bring too much computational burden.

Qualitative Results. We have visualized the comparison of different semantic segmentation methods and different components implementation strategies. Please refer to Sec. D in the supplementary material.

5 Conclusion

In this paper, we propose CACC framework to enhance model performance by leveraging internal consensus and conflicts cross data augmentations, eliminating the need for external components. The AFM module not only reduces the impact of unreliable but incorrect predictions by leveraging consensus in data augmentations, but also amplifies the low-confidence correct predictions. CTC module addresses the challenge of fixed thresholds by dynamically adapting them based on the confusion among categories. Extensive experiments across various benchmarks have demonstrated the superiority of the CACC, and ablation studies have further validated the effectiveness of each component.

Acknowledgments. This work is supported in part by Natural Science Foundation of Guangdong Province of China Under Grant No. 2024A1515011741, and partly supported by National Natural Science Foundation of China under Grant No. 62376292.

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
4. Ding, C., Zhang, J., Ding, H., Zhao, H., Wang, Z., Xing, T., Hu, R.: Decoupling with entropy-based equalization for semi-supervised semantic segmentation. In: IJCAI. pp. 663–671 (2023)
5. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman: The pascal visual object classes challenge: A retrospective. IJCV pp. 98–136 (2015)
6. Fan, S., Zhu, F., Feng, Z., Lv, Y., Song, M., Wang, F.Y.: Conservative-progressive collaborative learning for semi-supervised semantic segmentation. TIP (2023)
7. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV. pp. 991–998. IEEE (2011)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
9. Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., Wang, L.: Semi-supervised semantic segmentation via adaptive equalization learning. NeurIPS **34**, 22106–22118 (2021)
10. Jin, Y., Wang, J., Lin, D.: Semi-supervised semantic segmentation via gentle teaching assistant. NeurIPS **35**, 2803–2816 (2022)
11. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML. vol. 3, p. 896. Atlanta (2013)
12. Li, J., Dai, H., Han, H., Ding, Y.: Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In: CVPR. pp. 21694–21704 (2023)
13. Li, S., He, Y., Zhang, W., Zhang, W., Tan, X., Han, J., Ding, E., Wang, J.: Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. In: ICCV. pp. 16348–16358 (2023)

14. Liu, S., Zhi, S., Johns, E., Davison, A.: Bootstrapping semantic segmentation with regional contrast. In: ICLR (2021)
15. Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: CVPR. pp. 4258–4267 (2022)
16. Lv, X., Persello, C., Huang, X., Ming, D., Stein, A.: Deepmerge: Deep learning-based region-merging for image segmentation. arXiv preprint [arXiv:2305.19787](https://arxiv.org/abs/2305.19787) (2023)
17. Ma, J., Wang, C., Liu, Y., Lin, L., Li, G.: Enhanced soft label for semi-supervised semantic segmentation. In: ICCV. pp. 1185–1195 (2023)
18. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. *NeurIPS* **31** (2018)
19. Pan, W., Xu, Z., Yan, J., Wu, Z., Tong, R.K.y., Li, X., Yao, J.: Semi-supervised semantic segmentation meets masked modeling: Fine-grained locality learning matters in consistency regularization. arXiv preprint [arXiv:2312.08631](https://arxiv.org/abs/2312.08631) (2023)
20. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS* **33**, 596–608 (2020)
21. Sun, B., Yang, Y., Zhang, L., Cheng, M.M., Hou, Q.: Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. arXiv preprint [arXiv:2306.04300](https://arxiv.org/abs/2306.04300) (2023)
22. Wang, C., Xie, H., Yuan, Y., Fu, C., Yue, X.: Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation. In: ICCV. pp. 931–942 (2023)
23. Wang, H., Li, X.: Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In: MICCAI. pp. 582–591 (2023)
24. Wang, X., Zhang, B., Yu, L., Xiao, J.: Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. In: CVPR. pp. 3114–3123 (2023)
25. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: CVPR. pp. 4248–4257 (2022)
26. Wang, Z., Zhao, Z., Xing, X., Xu, D., Kong, X., Zhou, L.: Conflict-based cross-view consistency for semi-supervised semantic segmentation. In: CVPR. pp. 19585–19595 (2023)
27. Xie, H., Wang, C., Zheng, M., Dong, M., You, S., Fu, C., Xu, C.: Boosting semi-supervised semantic segmentation with probabilistic representations. In: AAAI. pp. 2938–2946 (2023)
28. Xu, H., Liu, L., Bian, Q., Yang, Z.: Semi-supervised semantic segmentation with prototype-based consistency regularization. *Adv. Neural. Inf. Process. Syst.* **35**, 26007–26020 (2022)
29. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: CVPR. pp. 7236–7246 (2023)
30. Yuan, J., Ge, J., Wang, Z., Liu, Y.: Semi-supervised semantic segmentation with mutual knowledge distillation. In: ACMML. pp. 5436–5444 (2023)
31. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV. pp. 6023–6032 (2019)

32. Zhao, Z., Long, S., Pi, J., Wang, J., Zhou, L.: Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In: CVPR. pp. 23705–23714 (2023)
33. 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: CVPR. pp. 11350–11359 (2023)



Event-Aware Multi-component (EMI) Loss for Fraud Detection

Tarun Somavarapu, Anand Vir Singh, Maneet Singh^(✉), Shraddha Pandey, Shantanu Verma, and Kushagra Agarwal

AI Garage, Mastercard, Gurgaon, India

{tarun.somavarapu, anandvirsingh.chauhan, maneet.singh}@mastercard.com

Abstract. Fraudulent transactions affect the different entities involved in the payment pipeline: (i) the merchant/vendor at which the transaction is performed, (ii) the authorizing bank, (iii) the card-holder, and (iv) the payment processing network/gateway. While fraud transactions result in the loss of billions of dollars globally, they may also result in reputational damage to the involved parties. Detecting fraudulent transactions is thus of utmost importance for the business and it also enhances customer experience in the financial domain. Predicting fraud transactions involves identifying suspicious transactions at the time of authorization and raising an alert to the decision-making authority. This research proposes a novel Event-aware Multi-component (*EMI*) loss for fraud prediction which incorporates key fraud-specific characteristics for learning a robust and accurate fraud prediction model. Specifically, the proposed loss incorporates key domain characteristics of fraud modeling such as focusing more on recent transactions, optimizing for an ideal event (fraud) rate, and maximizing the net benefit (or fraud savings) seen by the fraud prediction model. Further, the proposed loss is agnostic to the model architecture and can be utilized with different backbone architectures. Experimental results and analysis on multiple datasets demonstrate the efficacy of the proposed loss, where it achieves improved detection performance while optimizing for the above-mentioned industry requirements.

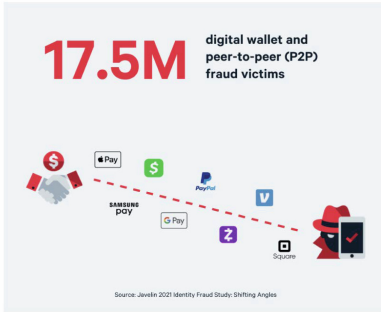
Keywords: Fraud Detection · Transaction Modelling · Financial Domain

1 Introduction

Fraudulent transaction prediction is one of the key challenges faced by the financial industry. Fraud transactions that are not blocked or detected in real-time often result in monetary loss, along with anguish to the card-holder (Fig. 1). Further, high fraud rates also affects the brand value/reputation of banks, payment networks, and merchants. As per a recent report in 2021, there was a

All authors were with AI Garage, Mastercard, India at the time of this research.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15327, pp. 105–119, 2025.
https://doi.org/10.1007/978-3-031-78398-2_7



(a) Large Scale Fraud Transactions in the Digital Domain



(b) Effect of Fraud Transactions on Customers

Fig. 1. (a) Fraudulent transaction attempts are growing in number, resulting in larger number of affected individuals and higher revenue loss. (b) Often, poor customer experience due to fraud transactions result in account closures at financial institutions. Detecting fraud transactions is thus an important task for all involved stakeholders. Both images have been taken from the Internet.

global loss of 20 billion US dollars due to fraudulent e-commerce transactions world-wide¹, which is a steep increase of 14% as compared to the previous year. Therefore, developing an efficient fraud prediction model is of utmost importance with large-scale real-world applicability and wide-scale impact. Transaction monitoring applications are often used by financial institutions such that they are able to monitor potentially risky transactions and take real-time decisions for fraud alerting. Often, in such scenarios the applications are required to provide a score for each transaction, following which the financial institutions might incorporate domain knowledge based business rules for ingesting the scores onto their final decision making pipeline. In order to eliminate the post-processing via business rules, this research proposes a novel model-agnostic Event-aware Multi-component loss for training a robust fraud prediction model, while incorporating key domain-specific characteristics during training.

The task of fraudulent transaction prediction involves taking real-time decisions on whether an incoming transaction is fraudulent or not [2, 9, 21, 23]. The task thus requires efficient algorithms (capable of running in real-time) while modeling key characteristics of fraudulent transactions in order to have high detection rates. In the literature, recent research has focused on proposing novel algorithms for the challenging problem of fraud transaction prediction. Given the sequential/time-based behavior of card-holders, most of the algorithms focus on modeling the previous history of the user using sequential models [11, 31]. Research has also focused on modeling the relationship between the different entities in the payment network (such as card-holder and merchant) using spa-

¹ <https://www.statista.com/statistics/1273177/ecommerce-payment-fraud-losses-globally/>.

tial/relational models [16]. While such algorithms focus on capturing the different patterns observed in the data (further elaborated in Section 2), to the best of our knowledge, they do not focus on incorporating the domain-specific business knowledge for generating robust fraud prediction models, suitable for applicability in the real world.

In order to address the above limitations, this research proposes a novel model-agnostic *Event-aware Multi-component (EMI)* loss for fraudulent transaction prediction. The proposed loss incorporates domain knowledge by modeling the transaction patterns while optimizing for the overall fraud predictions, net benefit (amount savings), and effective classification performance. Further, since fraud patterns often change over time, the proposed EMI loss focuses on providing higher importance to recent transactions in order to ensure recency based model learning. The model-agnostic property of the EMI loss enables its applicability across different architectures, thus also supporting quick real-time inference. To summarize, the key highlights of this research are as follows:

- A novel Event-aware Multi-component (EMI) loss function has been proposed for learning efficient fraud prediction models. The EMI loss utilizes key domain-specific knowledge for optimizing the number of fraud predictions over a batch of transactions. The proposed loss ensures that the model is deployable in real-world setups without hampering the customer experience (by not raising too many false alarms or false positives). Further, to the best of our knowledge, this is the first of its kind formulation which focuses on incorporating domain-specific business knowledge for fraud prediction models such as optimizing for the predicted event rate, net benefit, and recency based data importance.
- As shown in Fig. 2, the proposed EMI loss is model agnostic and can be applied with different backbone architectures (e.g., Multi Layer Perceptrons (MLPs), Long Short-Term Memory architectures (LSTMs), or transformers). Experimental analysis across different models demonstrates the model agnostic behavior of the loss, where enhanced performance is obtained with different backbone architectures.
- The efficacy of the proposed loss has been demonstrated on two financial datasets: (i) the Tabformer dataset [24] and (ii) an in-house synthetically generated dataset. Experimental analysis and results demonstrate the improvement obtained by the proposed EMI loss as compared to the baseline and other comparative techniques. Experiments have also been performed to further analyze the different components of the proposed loss via an ablation study, which strengthens the inclusion of different terms in the proposed formulation.

The remainder of this paper is organized as follows: Section 2 presents the related work in the area of automated fraud transaction prediction. Section 3 presents the proposed EMI loss, along with the detailed description of the framework. Section 4 presents the dataset details and protocols, followed by the results and analysis. Section 6 concludes the paper with the summary and future work.

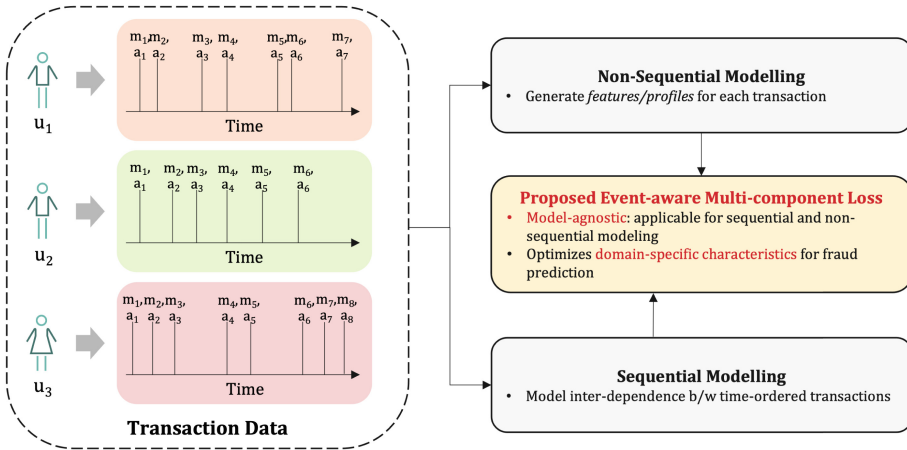


Fig. 2. In the literature, research has focused on developing sequential or non-sequential architectures for the given transaction data. The proposed loss can be applied to either kinds of deep learning models while also incorporating domain-specific characteristics for improved business relevance.

2 Related Work

The area of automated fraud transaction prediction has been an active area of research since the past several decades. Early fraud detection systems relied on rule-based techniques [20], wherein the system’s capabilities are limited by the experts’ knowledge [1]. Such engines typically hard-code business rules and thus suffer from limited real-world applicability. On the other hand, statistical methods try to learn these rules using the data by employing techniques like capturing interactions among features and generating summaries at transaction level [22, 34]. Association rules have also been used to detect credit card fraud [8] as well as other fraud types like healthcare fraud [28]. These methods try to infer IF-THEN-ELSE statements from the data itself, using symbolic/categorical features. For example, Brause *et al.* [8] proceed by comparing each transaction with another and finding pairs of similar ones (associations) and thus infer causality among sets of objects. Such raw associations lead to long rules, often diminishing the generalization ability of the model. Various techniques, like ignoring selective features for rule generation at varying tree levels, hope to shorten the rule length. Other statistical methods include clustering and outlier detection based on user analysis [6, 32] which fall under the unsupervised learning paradigm and do not utilize the rich label information available during model training.

In the literature, supervised learning-based algorithms have been proposed for fraud detection which either utilize the sequential information of a cardholder/user or rely on non-sequential cues for identifying fraudulent transactions. Non-sequential supervised learning algorithms rely on robust feature engineering in the form of *profiling*. Profiles are aggregations of data by a particular field,

or group of fields, over a pre-decided window of time to capture the historical behaviour and interactions among features and their interaction with fraud labels [22, 30]. Models like the Multi-Layer Perceptron (MLP), Decision Tree, or Support Vector Machine (SVM) often utilize a single transaction as input and thus are benefited by the feature engineering process, that would have otherwise been devoid of context. An example of such features can be the amount transacted by a user at a given merchant in the previous one hour, represented by *user_merch_amt_1hr*. Given a test transaction, this feature provides the model the ability to look back one hour, know whether the user was active/inactive, whether they are accustomed to having interactions with the merchant for the amount in question. In this way, such methods can identify relevant patterns in the scope of their features and respective time windows. Such heavy dependence on manual feature engineering limits the field of view of algorithms and thus limits their potential for real-world applications. This method of feature engineering is also time and compute expensive, and thus hard to utilize for quick inference requirements during run-time as well as challenging to engineer [7].

Sequential models aim to utilize the inherent ordering amongst subsequent transactions with respect to time by implicitly modeling this characteristic by design. Models like Convolutional Neural Networks (CNN) are known to capture short-term context, whereas others like Recurrent Neural Network (RNN) [4], Long-Short Term Memory Network (LSTM) [18], and gated-recurrent unit [13] can capture longer range dependencies as well. Zhang *et al.* [33] use convolution to capture interactions among subsequent features and extract the relevant derivative features, instead of interactions amongst subsequent transactions, for which they use a feature sequencing layer to learn the best order of features. In order to capture the longer range dependencies, Branco *et al.* [7] treat each entity/user as an independent sequence that share learnable parameters of the Gated Recurrent Unit (GRU). Such models focus on learning from the past events of the given user/card-holder. While these models don't rely on extensive feature engineering, it has been found that adding the manually engineered features appear to guide them further to an optimal solution [18]. Zhu *et al.* [36] captures relevant feature level and event level characteristics using field-level extractor and event-level extractor, respectively. While existing sequential and non-sequential learning-based algorithms have provided substantial advancements in the current literature, automated fraud transaction detection still remains to be a long standing problem [10, 12, 29, 31, 35]. We believe that a possible drawback of the existing solutions is the lack of domain/business specific knowledge inclusion during model development. Most of the current approaches aim to improve the modelling and representation of data from an architectural perspective, while ignoring the domain-specific business aspects of the same.

3 Proposed EMI Loss

This research focuses on bridging the above discussed gap and incorporating domain-specific insights for learning an efficient, robust, and deployable model.

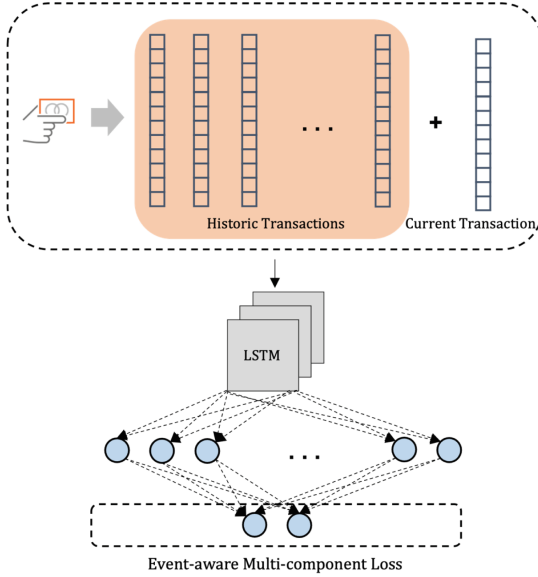


Fig. 3. Diagrammatic representation of the architecture used with the proposed EMI loss. The previous ten events (transactions) are provided to the LSTM model, followed by applying dense layers on the learned embedding. The error obtained via the proposed EMI loss is back-propagated throughout the network for effective model learning.

We aim to explicitly learn fraud prediction models that not only give better performance with respect to the various metrics but also improve the net benefit (fraud amount savings) incurred for direct business relevance while keeping an overall low fraud prediction alarms. From our domain understanding [14], we notice that recent trends in transactions are often more important than the older ones, and thus propose to incorporate the notion of *recency* during model training which can be useful for non-sequential networks [18] as well, thus leading to improved real-time impact.

For a given architecture, the loss function is the function that measures the distance between expected and predicted outputs. To create the best performing models, we are often required to minimize this function. In order to develop robust models, we should first identify important aspects needing to be optimized which can be included in the loss function. For example, while creating a fraud prediction model, the dollar amount saved (or the net benefit) [27] is often more important than the cross-entropy loss [15] values. These two are related but not always 100% correlated. Furthermore, the loss functions are mostly optimized using gradient descent [26] while training. Thus, they must be continuous and differentiable, this leads to achieving the global minimum loss values. To this effect, we propose a novel *Event-aware Multi-component (EMI)* loss function consisting of three components: (i) recency based cross-entropy loss ($\mathcal{L}_{Recency}$), (ii) predicted event rate optimization (\mathcal{L}_{PER}), and (iii) net benefit loss component

(\mathcal{L}_{Net}):

$$\mathcal{L}_{EMl} = \alpha \mathcal{L}_{Recency} + \beta \mathcal{L}_{PER} + \gamma \mathcal{L}_{Net} \quad (1)$$

where, α, β, γ correspond to the weight values given to each loss component. Details regarding each component are as follows:

(i) Recency based Cross-entropy Loss Component ($\mathcal{L}_{Recency}$): Transaction fraud monitoring can be referred to as a two class classification problem (fraud or non-fraud). As with traditional classification loss functions, the base architecture utilizes the standard Cross-Entropy loss for learning an effective classifier. The financial world often witnesses variations in the transaction patterns which are essential to model. The proposed recency based cross-entropy loss component introduces a *recency* factor during model training. The recency factor focuses on providing higher weight-age to recent transactions as compared to the older transactions, thus ensuring that the model is able to capture recent fraud trends. Mathematically, given a prediction y_i' for a target y_i , a recency weight of Δt_i is introduced:

$$\mathcal{L}_{Recency} = \sum \frac{y_i * \log y_i'}{\Delta t_i} \quad (2)$$

The recency weight is inversely proportional to the duration of the current transaction from the last transaction, thus resulting in higher weight-age to most recent transactions.

(ii) Predicted Event Rate (PER) Optimization Loss Component (\mathcal{L}_{PER}): The Predicted Event Rate (PER) is defined as count of fraud predictions over the count of total observations. On the other hand, the True Event Rate is defined as the count of the true frauds over the count of total observations. In an ideal scenario, the model should produce a PER close to true event rate, thus resulting in lesser false positives and promoting the model to predict the under-sampled class (fraud). In order to model the PER, the \mathcal{L}_{PER} has been formulated as follows:

$$\mathcal{L}_{PER} = (r - MSE(y_i', \mathbf{0}_n)) \quad (3)$$

where, r is the true event rate and $\mathbf{0}_n = (0, 0, \dots, 0) \in R_n$ is a zero vector. The above loss thus promotes the model's PER to be as close as possible to the true event rate.

(iii) Net Benefit Loss Component (\mathcal{L}_{Net}) [5]: This is one of the most important business metrics for the fraud prediction domain. When a fraud prediction model has been deployed by an entity, it would provide recommendations for blocking transactions that appear fraudulent to the model. For a true positive event (actual fraud), it will be saving the total sum of the amount of such transactions. On the other hand, for a false positive event (incorrect fraud prediction), it will lose out on the transaction processing fee that it would have gained by allowing the transaction through. So, effectively the net benefit (or fraud savings) can be viewed as the difference between the dollar amount of the true

positive transactions and the loss incurred due to the false positive transactions. For generality, we have defined the net benefit loss component as a function of the true positive transactions and the false positive transactions, which can be represented as follows:

$$\mathcal{L}_{Net} = f(\text{true positives}, \text{false positives}) \quad (4)$$

The proposed EMI loss is thus a combination of the above defined loss components: $\mathcal{L}_{Recency}$, \mathcal{L}_{PER} , \mathcal{L}_{Net} by using relevant hyper-parameters (α, β, γ) . Thus, the combined Event-aware Multi-component fraud prediction loss is defined as:

$$\mathcal{L}_{EMI}(y'_i, y_i) = \alpha * \sum \frac{y_i * \log y'_i}{\Delta t_i} + \beta * (r - MSE(y'_i, \mathbf{0}_n)) + \gamma * f(\text{true positives}, \text{false positives}) \quad (5)$$

The proposed Event-aware Multi-component loss function thus enables the learning of a robust fraud prediction model while incorporating the domain-specific business trends and requirements.

Table 1. Details regarding the datasets and protocols used in this research. Number of transactions (txns.) in the training, testing, and validation sets have been provided for ease in reproducibility. Both the datasets simulate real-world scenarios with a relatively smaller event rate (%) (fraud rate) across the transactions.

Dataset	Txns	Frauds	Event Rate	Users	Fields	Train	Val	Test
Tabformer [24]	24M	29757	0.122	6139	8	17M	2.4M	4.8M
In-house Synthetic	1M	67580	6.758	37404	8	700K	100K	200K

4 Experiments and Protocol

Experiments have been performed on two datasets with varying backbone architectures (LSTM, MLP, and RNN), along with detailed analysis of the EMI loss. The following subsections elaborate upon the dataset details and corresponding protocols, followed by the implementation details and the results.

4.1 Datasets and Protocol

The proposed EMI loss has been evaluated on two datasets containing transaction records. Table 1 presents the dataset details along with the training and testing protocols. The following paragraphs elaborate upon the protocols in detail:

- **Tabular Transformers for Modeling Multivariate Time Series Credit Card Dataset (Tabformer)** [24]: The dataset contains synthetically generated 24M credit card transactions from 20,000 users. The transactions have been generated using rule based generators, where the values are generated using stochastic sampling techniques, similar to the methods by Altman *et al.* [3]. Every transaction has 12 fields consisting of both continuous and discrete nominal attributes, such as the transaction amount, merchant location, transaction date etc. Similar to the existing protocol, we create samples by combining 10 contiguous rows (with a stride of 10) in a time-dependent manner for each user.
- **In-house Synthetic Dataset:** Owing to the limited availability of publicly available datasets containing fraud/non-fraud transaction information, experiments have also been performed on an in-house synthetic dataset. The dataset consists of 1M transactions from 37,500 unique users with 28,000 unique merchants and has a total of 67,580 fraudulent transactions. Every transaction consists of 298 unique fields like transaction time, merchant ID, etc. We create sequences of time, amount, and other fields in the input for each user, ordered by time and pass it then onto the model architecture. The complete dataset is divided into a training, validation, and testing partition, having 70%, 10%, and 20% of transactions, respectively in each set.

4.2 Implementation Details

For the two datasets, experiments have been performed with a LSTM [17] base architecture (Fig. 3) consisting of two hidden layers, followed by two dense (fully connected) layers for classification. The LSTM model has been implemented in the PyTorch environment [25] with a NVIDIA Quadro RTX6000 GPU. As demonstrated in Fig. 3, the past ten transactions of a user are provided to the LSTM model, followed by two dense layers for predicting whether the current transaction is fraudulent or not. Characteristics of the transactions are provided as input (depending upon the dataset and availability of information). Categorical features (e.g., merchant category code, industry, etc.) are converted into one-hot encoded embeddings, while numerical features (e.g., transaction amount) are provided as is. The predicted class scores are used in the loss equation Eq.5 and the weight parameters for the same are initialized as follows: $\alpha = 1e - 5$, $\beta = 1.5$ and $\gamma = 0.01$. The model is trained using the Adam optimizer [19] for 100 epochs with 1024 batch-size. Comparison has also been performed by replacing the LSTM with a RNN and a MLP architecture as well. The following section elaborates upon the results and analysis.

5 Results and Analysis

Tables 2-4 present the performance obtained by the EMI loss in different experiments. Comparison has been performed in terms of the precision, recall, F-1 score, and the net benefit obtained by the different models. Due to privacy concerns, we are unable to share the exact dollar savings (net benefit), and thus have

Table 2. Results obtained on the TabFormer dataset: precision, recall, F-1 score, and the percentage variation in the net benefit by each algorithm have been reported. The proposed EMI loss with a LSTM architecture demonstrates improved performance in terms of the overall F-1 score, while also obtaining a higher net benefit.

Algorithm	Precision	Recall	F-1 Score	AUC-PR	Benefit (%)
Multi-layer Perceptron	74.21	52.15	61.25	48.25	-67.2%
Recurrent Neural Network	89.02	75.46	81.68	81.49	-0.5%
Long Short-Term Memory	89.22	70.05	78.48	79.65	-6.5%
MLP + \mathcal{L}_{EMI}	86.11	51.55	64.49	49.51	-66.7%
RNN + \mathcal{L}_{EMI}	90.36	76.14	82.25	82.64	+2.6%
Proposed LSTM + \mathcal{L}_{EMI}	92.43	76.48	83.70	84.21	-

presented them in a relative format, that is, percentage increase or decrease. The key results are as follows:

Comparison on the TabFormer Dataset: As demonstrated in Table 2, the proposed EMI loss presents improved performance on the Tabformer dataset as compared to other techniques. Specifically, the proposed loss achieves a precision and recall of 92.43 and 76.48, respectively, demonstrating significant improvement from the LSTM model (trained only with the cross-entropy loss) which obtains a F-1 score of 78.48, which is around 5% lower than the proposed model (83.70). In the literature, a F-1 score of 86.00 has been reported on the Tabformer dataset [24], however, it is important to note that the authors follow a different protocol (exact splits are not available) and also perform additional steps of upsampling during training. Therefore, it is impossible to draw a fair comparison. Further, the existing research [24] presents an improvement of 3% from the then reported baseline, while we observe an improvement of around 5% from the base LSTM architecture. Comparison can also not be performed on the net benefit (in terms of the dollar amount) since that has not been reported in the existing research.

Table 3. Ablation study of the proposed EMI loss on the TabFormer dataset. The experiments demonstrate contribution of each loss component towards the creation of a robust fraud prediction model. Benefit (%) presents the percentage variation in the net amount benefit (in dollars) by each model.

Algorithm	Precision	Recall	F-1 Score	Benefit (%)
$\mathcal{L}_{EMI} - \mathcal{L}_{Recency}$	94.46	72.25	81.87	+0.7%
$\mathcal{L}_{EMI} - \mathcal{L}_{Net}$	91.64	74.28	82.05	-2.9%
$\mathcal{L}_{EMI} - \mathcal{L}_{PER}$	92.48	72.92	81.55	-5.1%
Proposed \mathcal{L}_{EMI}	92.43	76.48	83.70	-

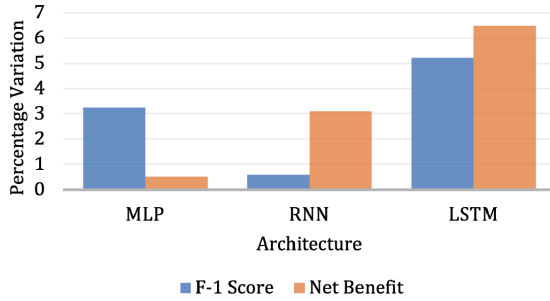


Fig. 4. Improvement in the F-1 score and net benefit (percentage) on training the three architectures with the proposed EMI loss. Maximum improvement is seen with the LSTM model, possibly due to their better capabilities of modeling long term sequential data.

Effect of Different Backbone Architectures: The efficacy of the proposed loss has also been demonstrated by using different backbone architectures (RNN, MLP, LSTM), where, training with the proposed EMI loss demonstrates improvement as compared to training with the native loss function. Table 2 presents the performance comparison on different backbone architectures with and without the proposed loss. The improved performance (in terms of the precision/recall/F-1 score) thus supports the model-agnostic behavior of the proposed loss. For example, an improvement of over 3% is observed upon utilizing the EMI loss with the LSTM architecture as compared to the native loss (92.43 versus 89.22). The benefit of using sequential models is also visible across different architectures, where a difference of at least 15% in precision is observed between the MLP architecture and the RNN/LSTM architectures. Further, it is also interesting to note that MLP based architectures appear to perform substantially poorer in terms of modelling the effective net benefit. Fig. 4 also presents a bar graph showcasing the improvement in the F-1 score and the net benefit percentage on training the three architectures with the proposed loss as compared to the native cross-entropy loss. Maximum improvement is observed with the LSTM architecture for both the metrics. Overall, the best performance is obtained by utilizing the LSTM architecture with the proposed EMI loss in terms of the accuracy metrics (precision/recall), while a minor trade-off is seen in the overall net benefit as compared to the RNN architecture with the EMI loss.

Ablation Study on the EMI Loss: Table 3 presents the ablation study on the proposed EMI loss using the TabFormer dataset. Experiments have been performed to understand the contribution of each loss component by training the LSTM architecture with the proposed loss function after removing each term and analyzing the performance metrics. As can be observed from Table 3, removal of any component from the EMI loss results in a drop in performance in terms of the precision, recall, and F-1 scores. Maximum drop in performance is observed upon removing the PER loss component (\mathcal{L}_{PER}), resulting in a drop in F-1 score from 83.70 to 81.55. The performance drop appears intuitive in

nature since the PER loss component controls the total fraud predictions, thus pushing the model towards predicting lesser false positives. On the other hand, minimal impact is seen upon removing the recency based component ($\mathcal{L}_{Recency}$), where a drop in F-1 score is accompanied with a slight increase (0.7%) in the net benefit. Overall, the ablation study supports the inclusion of the different loss components for achieving enhanced fraud prediction performance and business objectives.

Table 4. Results obtained on the in-house synthetic dataset. The proposed technique demonstrates an improvement in the overall F-1 score and a 3.5% increase in the net benefit.

Algorithm	Precision	Recall	F-1 Score	Benefit (%)
LSTM	33.80	49.81	40.31	-3.5%
Proposed	39.76	48.69	43.75	-

Performance on the In-house Synthetic Dataset: Table 4 presents the performance comparison between the proposed loss and the baseline model on the in-house synthetic dataset. An increase in the F1-Score is observed from 40.31 to 43.75, along with an increase in the precision (from 33.80 to 39.76) for the fraud class when using the proposed loss. An improvement of +3.5% is also observed for the overall net benefit.

The above set of experiments and results suggest improved performance by the proposed Event-aware Multi-component (EM) loss function. Improvement can be seen across capturing fraudulent transactions and the overall net benefit (in terms of dollar amount). The improved performance strengthens its usage for real-time fraud prediction, while providing flexibility during model training for tuning different business aspects.

6 Conclusion and Discussion

The past few decades have witnessed tremendous growth in the domain of Data Mining, Machine Learning, and Artificial Intelligence. While initial research began with modelling a handful of data points, current research focuses on understanding patterns and trends across millions of instances. The research developments have further enabled the adoption of such techniques/models in the real-world industrial setups as well. These days, it is not uncommon to automate mundane tasks and rely on learned AI models for day-to-day predictions such as weather forecasts, market trends, transaction monitoring, entertainment recommendations (dining/content consumption), or customer support for troubleshooting commonly faced issues. The omnipresence of such algorithms has thus enabled widespread utility of AI based services to the consumers (individuals) as well as institutions/corporations.

The financial world is one such domain consisting of billions of data points (transactions) and being fed by multiple automated services for several tasks. For example, automated data cleansing including noise removal, credit score assessment, market analysis, transaction risk monitoring, financial (credit/debit) limit adjustments, loan approvals, etc. The wide applicability of financial solutions across different demographics and geographies makes it an imperative domain for deployment of automated services. Despite the multiple applications and solutions, fraudulent transaction detection remains a long-standing challenging problem requiring dedicated research focus. From the real-world data, it is evident that fraudulent transactions are becoming more and more difficult to detect even with state-of-the-art algorithms developed in the past five years. Further, newer methods of transactions bring with them more innovative ways in which criminals can commit fraud. Thus, in order to keep up with the emerging trends in today's modern world, it is imperative to keep researching and developing more ingenious solutions to this problem.

To this effect, this research proposes a novel Event-aware Multi-component (EMI) loss function which incorporates domain-specific knowledge for building robust fraud prediction models. The proposed loss can be used for training a transaction monitoring model which provides a score of *riskiness* for each transaction, thus determining whether it is fraudulent or not. The EMI loss focuses on increasing the net benefit (in terms of the fraud amount savings) which is often a function of the true positives, false positives, and false negatives; minimizing the total fraud predictions (in order to reduce false positives) such that the model provides confident predictions; and gives higher importance to recent transactions (recency based learning). Experimental results and analysis across two different datasets (one open source and one in-house synthetic dataset generated from modelling the real-world transaction distribution) demonstrate the effectiveness of the proposed loss. Further, the inclusion of business-specific loss components allows the model to be deployable, while supporting a smaller model for quick real-time inference.

Despite the research advancements, we believe that there is still a long way to go in the field of fraudulent transaction detection. Novel techniques when creatively implemented to solve this problem can potentially help in creating better and more robust models. As part of future work, our efforts will be focused on incorporating better backbone architectures with the proposed EMI loss for robust model creation. Further, improvements can also be made at the input-level, where novel features can be identified and provided as input to the model for developing further enhanced models.

References


1. Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: A survey. *J. Netw. Comput. Appl.* **68**, 90–113 (2016). <https://doi.org/10.1016/j.jnca.2016.04.007>
2. AbdulSattar, K., Hammad, M.: Fraudulent transaction detection in fintech using machine learning algorithms. In: *International Conference on Innovation and Intelligence for Informatics, Computing and Technologies*. pp. 1–6 (2020)

3. Altman, E.: Synthesizing credit card transactions. In: ACM International Conference on AI in Finance. pp. 1–9 (2021)
4. Babaev, D., Savchenko, M., Tuzhilin, A., Umerenkov, D.: ET-RNN: Applying deep learning to credit loan applications. In: SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2183–2190 (2019)
5. Beneish, M.D., Vorst, P.: The cost of fraud prediction errors. Kelley School of Business Research Paper (2020-55) (2021)
6. Bolton, R.J., Hand, D.J.: Statistical fraud detection: A review. *Stat. Sci.* **17**(3), 235–255 (2002)
7. Branco, B., Abreu, P., Gomes, A.S., Almeida, M.S., Ascensão, J.T., Bizarro, P.: Interleaved sequence rnns for fraud detection. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3101–3109 (2020)
8. Brause, R., Langsdorf, T., Hepp, M.: Neural data mining for credit card fraud detection. In: International Conference on Tools with Artificial Intelligence. pp. 103–106 (1999)
9. Chen, J.I.Z., Lai, K.L.: Deep convolution neural network model for credit-card fraud detection and alert. *J. Artif. Intell.* **3**(02), 101–112 (2021)
10. Chen, L., Zhang, Z., Liu, Q., Yang, L., Meng, Y., Wang, P.: A method for online transaction fraud detection based on individual behavior. In: ACM Turing Celebration Conference - China (2019)
11. Cheng, D., Wang, X., Zhang, Y., Zhang, L.: Graph neural network for fraud detection via spatial-temporal attention. *IEEE Trans. Knowl. Data Eng.* **34**(8), 3800–3813 (2020)
12. Cheng, D., Xiang, S., Shang, C., Zhang, Y., Yang, F., Zhang, L.: Spatio-temporal attention-based neural network for credit card fraud detection. *AAAI Conference on Artificial Intelligence* pp. 362–369 (2020)
13. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint [arXiv:1412.3555](https://arxiv.org/abs/1412.3555) (2014)
14. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.A., Waterschoot, S., Bontempi, G.: Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Syst. Appl.* **41**(10), 4915–4928 (2014)
15. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134**(1), 19–67 (2005)
16. Gramopadhye, M., Singh, S., Agarwal, K., Srivasatava, N., Singh, A.M., Asthana, S., Arora, A.: CuRL: Coupled representation learning of cards and merchants to detect transaction frauds. In: International Conference on Artificial Neural Networks. pp. 16–29 (2021)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
18. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.E., He-Guelton, L., Caelen, O.: Sequence classification for credit-card fraud detection. *Expert Syst. Appl.* **100**, 234–245 (2018)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
20. Li, J., Huang, K., Jin, J., Shi, J.: A survey on statistical methods for health care fraud detection. *Health Care Manag. Sci.* **11**(3), 275–287 (2008). <https://doi.org/10.1007/s10729-007-9045-4>
21. Lim, K.S., Lee, L.H., Sim, Y.W.: A review of machine learning algorithms for fraud detection in credit card transaction. *International Journal of Computer Science & Network Security* **21**(9), 31–40 (2021)

22. Lin, W., Sun, L., Zhong, Q., Liu, C., Feng, J., Ao, X., Yang, H.: Online credit payment fraud detection via structure-aware hierarchical recurrent neural network. In: International Joint Conference on Artificial Intelligence. pp. 3670–3676 (2021). 10.24963/ijcai.2021/505
23. Lucas, Y., Jurgovsky, J.: Credit card fraud detection using machine learning: A survey. arXiv preprint [arXiv:2010.06479](https://arxiv.org/abs/2010.06479) (2020)
24. Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., Altman, E.: Tabular transformers for modeling multivariate time series. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3565–3569 (2021)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035 (2019)
26. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint [arXiv:1609.04747](https://arxiv.org/abs/1609.04747) (2016)
27. Stolfo, S.J., Fan, W., Lee, W., Prodromidis, A., Chan, P.K.: Cost-based modeling for fraud and intrusion detection: Results from the jam project. In: Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00. vol. 2, pp. 130–144 (2000)
28. Viveros, M.S., Nearhos, J.P., Rothman, M.J.: Applying data mining techniques to a health insurance information system. In: International Conference on Very Large Data Bases. p. 286–294. Morgan Kaufmann Publishers Inc. (1996)
29. Wang, Z., Wang, F., Zhang, H., Yang, M., Cao, S., Wen, Z., Zhang, Z.: 'Could You Describe the Reason for the Transfer?': A Reinforcement Learning Based Voice-Enabled Bot Protecting Customers from Financial Frauds, p. 4214–4223 (2021)
30. Xie, Y., Liu, G., Cao, R., Li, Z., Yan, C., Jiang, C.: A feature extraction method for credit card fraud detection. In: International Conference on Intelligent Autonomous Systems. pp. 70–75 (2019)
31. Zhang, C., Wang, Q., Liu, T., Lu, X., Hong, J., Han, B., Gong, C.: Fraud Detection under Multi-Sourced Extremely Noisy Annotations, p. 2497–2506 (2021)
32. Zhang, Z., Chen, L., Liu, Q., Wang, P.: A fraud detection method for low-frequency transaction. *IEEE Access* **8**, 25210–25220 (2020)
33. Zhang, Z., Zhou, X., Zhang, X., Wang, L., Wang, P.: A model based on convolutional neural network for online transaction fraud detection. *Security and Communication Networks* (2018)
34. Zheng, L., Liu, G., Yan, C., Jiang, C.: Transaction fraud detection based on total order relation and behavior diversity. *IEEE Transactions on Computational Social Systems* **5**(3), 796–806 (2018)
35. Zheng, L., Liu, G., Yan, C., Jiang, C., Zhou, M., Li, M.: Improved tradaboost and its application to transaction fraud detection. *IEEE Transactions on Computational Social Systems* **7**(5), 1304–1316 (2020)
36. Zhu, Y., Xi, D., Song, B., Zhuang, F., Chen, S., Gu, X., He, Q.: Modeling users' behavior sequences with hierarchical explainable network for cross-domain fraud detection. In: Proceedings of The Web Conference. pp. 928–938 (2020)



A Simple Heuristic for Controlling Human Workload in Learning to Defer

Andrew Ponomarev^(✉) 

St. Petersburg Federal Research Center of the Russian Academy of Sciences,
14th Line 39, 199178 St. Petersburg, Russia
ponomarev@iiias.spb.su

Abstract. In many cases, machine learning model is used not autonomously, but as a part of some larger system that may include human experts. Learning to defer technique allows to train models that can take into account error probabilities of both machine learning model and human expert and route samples accordingly in order to maximize overall accuracy of the system. However, most of the learning to defer methods don't allow constraining the deferral fraction, which is important, as the number of human experts and their capacity are usually limited. The paper proposes and explores a simple yet effective heuristic technique allowing to impose constraints on the fraction of samples deferred to an expert, thereby, helping to balance accuracy and coverage metrics. The technique can be used in conjunction with many existing learning to defer and rejection learning methods; it is evaluated using three popular learning to defer techniques and two datasets — a synthetic and a real-life, collected using crowdsourcing.

Keywords: Learning to defer · Human-AI complementarity · Human-AI collaboration

1 Introduction

Machine learning models taking into account human-AI collaboration, e.g., simply by avoiding to classify samples for which they are uncertain are especially demanded in responsible and critical applications. The most widely used technique to build such models is to train a model (e.g., for the classification) and then, during inference, estimate the uncertainty of the prediction and redirect to the human expert only those samples, for which the model is uncertain. There are several approaches for estimating uncertainty [3, 6, 9], potentially, any of them can be used. The main drawback of this approach is that it ignores the limited knowledge of the human (and respective probability of an error), considering him/her as an oracle.

To account for the limited human knowledge, a *learning to defer* paradigm has been proposed [11, 14]. The idea is to train two models — main model and a

The research is funded by the Russian Science Foundation (project 24-21-00337).

rejector model. The first one is responsible for solving the classification problem *per se*, while the second one decides which samples should be assigned to the main model, and which to the human. These models can be trained jointly or separately [2], but their training typically relies on some specially constructed loss function, taking into account human errors in the training sample. Therefore, the models learn also the distribution of human expertise in the feature space and act accordingly. However, a severe limitation of this approach is that it doesn't account for possibly limited human resources, besides, some of the methods of this group rely on the relative weight of terms in the loss function which can be hard to set up correctly.

As it is shown in [1,13], *learning to defer* problem with limited human resources can be formulated as a *mixed-integer linear programming problem* (MILP), where coverage can be just one of the constraints (e.g., [13] introduce also fairness constraints). While the solution of such MILP can provide an optimal sample allocation between human and model (in some sense and with certain assumptions), this setting has two potential drawbacks: a) these algorithms are designed to distribute a given set of samples, so they may be less convenient for a situation when the samples need to be distributed as soon as they arrive, b) solving a MILP problem can be computationally expensive.

This paper proposes and explores a simple yet effective heuristic technique to account for the situation when human resources are limited, allowing to balance accuracy and coverage metrics while distributing samples “on the fly”. The proposed heuristic is not intended to replace the existing learning to defer approaches (based on unconstrained optimisation of surrogate loss functions), instead, it aims to complement them.

The rest of the paper is structured as follows. Section 2 describes related publications. Section 3 describes the formal problem definition and the proposed technique. Section 4 contains the results of the evaluation of the proposed technique using several datasets.

2 Related Work

The problem of joint work of an AI model and a human expert has been attracting attention of the ML researchers for a relatively long time. Probably, the most developed setting is so-called *rejection learning*, when a model is trained to reject certain samples to improve reliability on the others [7]. However, in rejection learning human expert is typically out of the scope (or considered to be absolutely accurate).

More recently, a concept of *learning to defer* has been proposed [11], where human expert has been placed “into” the system, which means that these algorithms optimize deferral policies taking into account not only the performance of the machine learning model in various regions of the feature space, but also accuracy of the human expert which may also be different in the different regions of the feature space.

Most approaches to learning to defer rely on specially constructed loss functions, balancing components responsible for automated classifier and human

expert. These loss functions can either be very closely based on the utility theory [17] or further elaborated as surrogate loss functions with better optimization and calibration properties, see, e.g. [12, 14, 16].

However, all of these approaches are either aimed to find a deferral policy optimizing only the system accuracy, or rely on relative weighting in the loss function, which should reflect relative cost of sample processing by the model and the human expert, however, may be hard to specify. Crucial is that most of these approaches ignore that the expert capacity can be limited [10].

A few methods have been proposed to deal with the limited expert capacity, typically via MILP reformulation of learning to defer problem, e.g. [1, 4, 5, 13]. In [4, 5], MILP is solved for the training set and then its solution is approximated by an additional model (the proposed MILP formulation is limited to certain types of models). In [1, 13] MILP is built during inference, therefore, these approaches can only allocate a set of samples, but are less suited for the situation when the samples have to be distributed between model and human expert as soon as they arrive. Besides, solving MILP can be computationally expensive, especially for large datasets.

This paper explores a heuristic method for *on the fly* sample allocation respecting limited expert capacity, which can be used with existing learning to defer approaches and doesn't require solving MILP.

3 The Proposed Technique

3.1 Problem Definition

Given the dataset $\{(X_i, m_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, where $X_i \in \mathcal{X}$ are the features describing objects, $y_i \in \mathcal{Y}$ are true labels, and $m_i \in \mathcal{Y}$ are expert labels (not necessarily equal to the true labels, as the expert can also be wrong). The problem is find two functions — a classifier $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a *rejector* $r: \mathcal{X} \rightarrow \{0, 1\}$. To obtain a decision \hat{y}_i for some instance x_i these functions are combined in the following way:

$$\hat{y}_i = \begin{cases} h(x_i), & \text{if } r(x_i) = 1 \\ m_i, & \text{if } r(x_i) = 0 \end{cases} \quad (1)$$

Moreover, as access to the expert evaluations can be limited in the inference time, these functions must maximize classification quality, respecting the restriction on the number of used expert evaluations. Formally,

$$(h^*, r^*) = \arg \max_{(h, r)} \mathbb{E}_{(x, m, y) \in \mathcal{D}} [\mathbb{I}\{r(x)h(x) + (1 - r(x))m\} = y], \text{ s.t. } \mathbb{E}[r(x)] \geq C_r. \quad (2)$$

In practise, expectations in the equation above are typically estimated via empirical metrics, evaluated using a test dataset, sampled from the same distribution \mathcal{D} . Expectation of the number of correct answers corresponds to accuracy and the expectation of the samples assigned to the model h corresponds to coverage (ratio of the samples classified by the model, without resorting to the expert).

3.2 Loss Functions

The requirement to account for errors of both a machine learning model and an expert translates to natural loss function used to train h and r [11]:

$$\mathcal{L}_{nat}(x_i, m_i, y_i, h, r) = r(x_i)\ell_m(h(x_i), y_i) + (1 - r(x_i))\ell_{exp}(m_i, y_i),$$

where ℓ_m is a model loss and ℓ_{exp} is an expert loss. For example, if it is a binary classification problem, then ℓ_m can be a binary cross entropy (for the ℓ_{exp} it is a bit more complicated, see below).

However, direct using of such natural loss has two main drawbacks:

1. In the training data, m represents the class provided by the end user, therefore, if the expert is wrong, binary cross entropy becomes infinite. In practice, one can either clamp too large values (e.g., it is a default work-around for infinite log implemented in PyTorch) or use some other loss function (e.g., \mathcal{L}_1). In either case, a particular value, corresponding to the user error, must be somehow scaled to the range of the first term (ℓ_m).
2. The loss function optimizes the models only in terms of accuracy, not paying attention to the fact, that access to the pool of human experts might be limited.

Some surrogate losses has been proposed in the literature (e.g., [14]), mitigating the first problem, but they still optimize only in terms of accuracy.

3.3 Method

The proposed technique relies on a possibility to obtain a score, reflecting relative confidence of the classification of the sample by the model w.r.t. the classification of the same sample by the human expert. We will denote this score as $score_{(h,r)}(x)$. Absolute values of this score don't matter, instead, this score establishes an ordering: if $score_{(h,r)}(x_i) > score_{(h,r)}(x_j)$, then assigning x_i to the model (rather than human expert) will result in less probability of an error, than assigning x_j to the model. We provide specific examples of building $score_{(h,r)}(x)$ for several existing learning to defer methods later in this section.

The technique consists of three steps:

1. Train h and r models using an existing learning to defer algorithm. The resulting pair can perform deferral, typically achieving good (or, optimal in some sense) accuracy, but not respecting coverage constraints.
2. Using separate dataset $\mathcal{V}^c = \{(X_i^c, m_i^c, y_i^c)\} \sim \mathcal{D}$ and Algorithm 1, find the score threshold θ , corresponding to the required coverage C_r .
3. Apply Algorithm 2 to classify any incoming instances, sampled from the \mathcal{D} .

Training algorithm (Algorithm 1) evaluates scores for all the samples of \mathcal{V}^c and then considers each score value as a candidate threshold for assigning all samples with greater (or equal) scores to the model and the rest to the human expert (which is done using function f_{s_i} , defined as $h(x_j)$ if $score_{(h,r)}(x_j) \geq s_i$

and m_j otherwise). It then evaluates accuracy and coverage of each such split and picks a score value, such that: at least C_r (the required coverage) samples of \mathcal{V}^c have greater scores and the accuracy is maximal (among all the values, respecting the required coverage constraint).

Algorithm 1. Training

Require: $C_r \in [0; 1], \mathcal{V}^c, h, r$

Ensure: θ

```

for all  $i \in |X^c|$  do
   $s_i \leftarrow \text{score}_{(h,r)}(x_i)$ 
end for
 $a \leftarrow 0$ 
 $\theta \leftarrow \text{None}$ 
for all  $i \in |X^c|$  do
   $\tilde{a} \leftarrow \text{Accuracy}(f_{s_i}(X^c, m^c), y^c)$ 
   $\tilde{c} \leftarrow |\{j | j \in \{1, \dots, |X^c|\}, s_j \geq s_i\}| / |X^c|$ 
  if  $\tilde{c} \geq C_r$  and  $\tilde{a} > a$  then
     $a \leftarrow \tilde{a}$ 
     $\theta \leftarrow s_i$ 
  end if
end for

```

Inference algorithm (Algorithm 2) evaluates score of the passed instance, compares it with the threshold, found during training and redirects it accordingly (ASK). Note, that the algorithm uses only the value of the parameter θ , estimated using Algorithm 1, and the sample x , therefore, it can be applied to the samples as soon as they arrive, without the need for a large batch (unlike, e.g. [1]).

Algorithm 2. Inference

Require: x, θ

```

if  $\text{score}_{(h,r)}(x) \geq \theta$  then
  return  $h(x)$ 
else
  return ASK
end if

```

Lets consider scoring functions for several learning to defer algorithms. The simplest algorithm is confidence threshold-based (we'll refer to it as **Threshold**). It doesn't have a separate rejection model, but redirects to human expert instances, for which maximal softmax output [3] is lower than certain threshold (threshold is typically set to maximize accuracy). This algorithm is inherently not sensitive to the expertise variability in the input domain, so $\text{score}_{(h,r)}(x)$ can be defined as any confidence measure, e.g., maximal softmax output.

Another approach is to use natural loss directly, training two separate functions — classifier h and rejector r (**NatLoss**). In this case, there is a separate rejector, which output (sigmoid function, before binarization) gives the value appropriate for $score_{(h,r)}(x)$.

Finally, softmax parametrization loss, proposed in [14] (SP). In this case, h and r are modeled using using one neural model with $K + 1$ outputs (where K is the number of classes). Outputs 1 to K correspond to class probabilities, and $(K + 1)$ -th output correspond to deferral to the human expert. Let $\hat{p}_i(x)$ be the value of the i -th output. In this model, $h(x)$ is defined as $\arg \max_{k \in 1:K} \hat{p}_k(x)$, and $r(x)$ is 1 iff $\max_{k \in 1:K} \hat{p}_k(x) \geq \hat{p}_{K+1}(x)$ and 0 otherwise. The $score_{(h,r)}(x)$ can be defined as $\max_{k \in 1:K} \hat{p}_k(x) - \hat{p}_{K+1}(x)$. Intuitively, it reflects the difference between classifier’s confidence and estimated human expert’s reliability for the considered sample.

4 Evaluation

4.1 Datasets

The approach is illustrated using two datasets: synthetic and real-life. Synthetic dataset is designed for binary classification and is generated using following equations. For class ($y = 0$):

$$x_2 \sim U(0, 1) \tag{3}$$

$$x_1 \sim N(2x_2, \sigma) \tag{4}$$

For class ($y = 1$):

$$x_2 \sim U(0, 1) \tag{5}$$

$$x_1 \sim N(2 - 2x_2, \sigma) \tag{6}$$

And expert labels are defined by the following function:

$$m(x_1, x_2, y) = \begin{cases} y, & \text{if } \text{Bernoulli}(x_2) \\ 1 - y, & \text{otherwise} \end{cases} \tag{7}$$

A sample from this dataset is shown in Fig. 1. The plot on the left shows two classes and the plot on the right uses color to highlight instances for which the expert gives correct (green) and incorrect (red) result. The idea is that there is a region where the classes are separated relatively well (low values of x_2), and a region, where they are hard to separate (higher values of x_2 , at the same time, expert’s competence is highest for the high values of x_2 and lowest for the low values. As a result, it can be expected, that for the low values of x_2 the classification would be done by the model, and for the high values — delegated to the expert. The σ parameter of the normal distribution controls the overlap

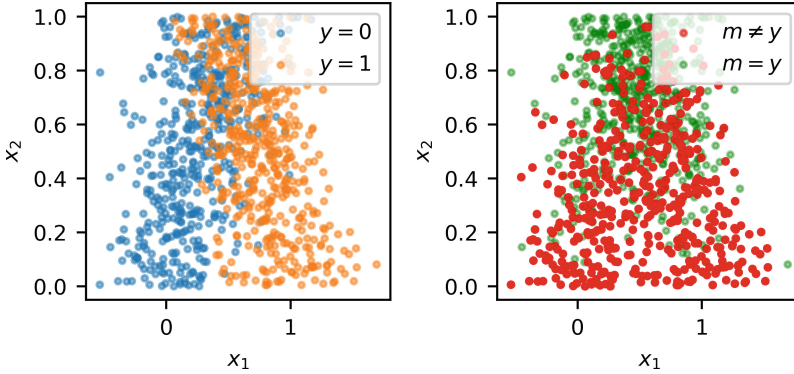


Fig. 1. A sample from the synthetic dataset.

between classes and limits the classification model accuracy (even for the low values of x_2). In the experiments, presented in this paper, the value of σ is set to 0.25 (the same value is used in the sample shown in Fig. 1).

Real-life dataset is CIFAR-10H [15], a subset of CIFAR-10 [8], containing 10,000 images, for which human labels were collected using crowdsourcing. There are several human labels for each image in this dataset, and to produce m_i one of them was picked randomly.

4.2 Models

For the synthetic dataset we used a multi-layer perceptron (MLP) architecture with two hidden layers (each with 40 neurons) and ReLU activation. The the number of such models, output layer configuration and loss function were different for the examined *learning to defer* approaches:

- for **Threshold** it was one MLP model with two output neurons (corresponding to classes) trained using cross entropy using only ground truth labels;
- for **NatLoss** it were two MLP models (one for h and one for r), each with one output neuron, trained simultaneously using \mathcal{L}_{nat} ;
- for **SP** it was one MLP model with three output neurons (two of them correspond to classes and one for the deferral to the expert) trained using softmax parametrization loss from [14].

In all cases, the optimizer was Adam, and training was done in batch mode until convergence (training set loss change less than 10^{-5}).

For the CIFAR-10H we trained a ResNet-18 model on the CIFAR-10 dataset (excluding the images that are also part of the CIFAR-10H) to obtain classification accuracy of about 86%. Then we transformed CIFAR-10H images into their compressed representations of width 512, taking the output of the layer just before the classification head. All the models for human-AI learning (classification and deferral policies) are MLPs with two hidden layers of 80 and 40

neurons and ReLU activation function. Output layer configuration and loss functions were the same as described above (but with 10 classes).

4.3 Results

The main question that has to be answered during evaluation is if the heuristic technique described in Sec. 3 allows one to impose coverage constraint during inference time. The behaviour of one particular model w.r.t this requirement can be visualized using required coverage vs. test coverage plots (or, CC-plots). On the x-axis is required coverage (set during model training), on the y-axis is the test coverage evaluated using the test set. Note, that according to formal definition of the problem (Sec. 3), test coverage must be greater or equal to the required coverage, therefore, the graph should be above the diagonal line.

Fig. 2 shows example CC-plots for the synthetic dataset (left) and for CIFAR-10H dataset (right). It can be seen, that test coverage actually follows the required coverage in certain required coverage range, but for low required coverage, test coverage turns out to be significantly greater than required. This happens because algorithm in Sec. 3 looks for the parameter value maximizing accuracy and respecting coverage constraint. However, for certain coverage, maximum accuracy is obtained, therefore, for all required coverage values lower than this one, the same test coverage is returned (corresponding to the overall best accuracy of the human-AI system).

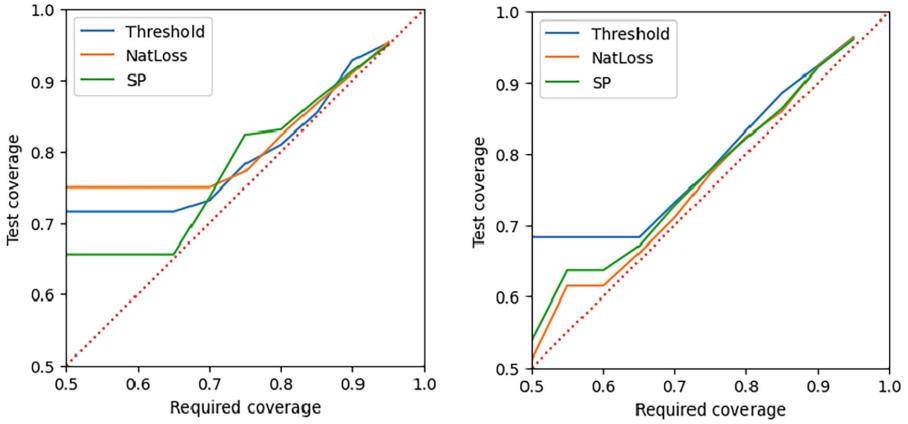


Fig. 2. Required coverage vs. test coverage for the synthetic dataset (left) and CIFAR-10H (right).

CC-plots can give insights to the behavior of the algorithm using certain trained model and certain human labels. However, it lacks generalization. For this purpose, we also build CC violation plots, showing the distribution of maximum violation of the coverage requirement for model. Fig. 3 shows an example of such

plot, built for 100 generated synthetic datasets (and respective models). We can see, that coverage constraint can be violated, but in vast majority of cases this violation is less than 1% and in all the cases it is less than 5%.

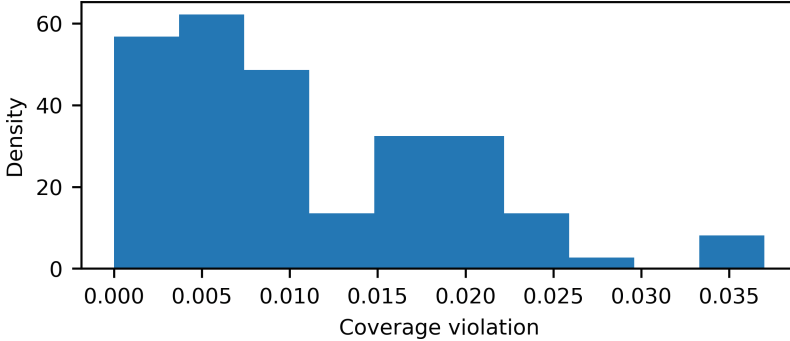


Fig. 3. Distribution of maximal coverage violation.

Finally, Fig. 4 shows examples of coverage-accuracy graphs induced by the proposed algorithm. In the scope of this paper, we do not consider the problem of selecting the best training algorithm (Threshold, NatLoss, SP or something else), rather, we aim to show that the simple technique described in Sec. 3 allows for adding coverage constraint to any of them. Coverage-accuracy graphs in this context are secondary, but their examination explains the patterns found on CC-plots (coverage with overall high accuracy).

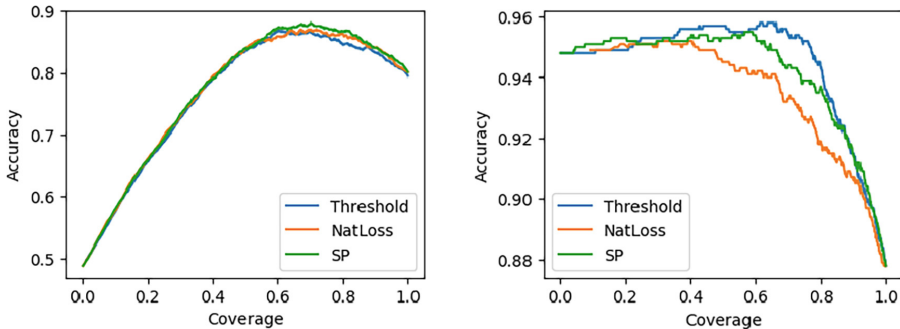


Fig. 4. Coverage vs. accuracy plots for the synthetic dataset (left) and CIFAR-10H (right).

An interesting observation is that coverage-accuracy graphs in the considered examples are unimodal curves. The ultimate goal of the algorithm 1 is to find maximum accuracy in certain coverage region, therefore, the unimodality of the

coverage-accuracy gives an opportunity to significantly improve time complexity of this algorithm by employing zero-order unidirectional optimization algorithms (e.g., golden section search). However, this effect requires further research and proper theoretical grounding.

5 Conclusion

The paper proposes and analyzes empirically a simple heuristic technique to help imposing coverage constraints on existing learning to defer models. It has been evaluated in conjunction with three popular deferral techniques — confidence-based, natural deferral loss-based and the one based on a surrogate loss function. Computational experiments using two datasets — a synthetic one and CIFAR-10H dataset collected using crowdsourcing — has shown that the imposed coverage constraint is respected during inference time, violations are possible, but they are not very numerous and not severe.

Future work is mostly connected to improving the complexity of the algorithm (accuracy-coverage curves give one possible idea for that), and to estimating theoretical guarantees of the constraint violation probability.

Acknowledgements. The research is funded by the Russian Science Foundation (project 24-21-00337).


References

1. Alves, J.V., Leitão, D., Jesus, S., Sampaio, M.O.P., Liébana, J., Saleiro, P., Figueiredo, M.A.T., Bizarro, P.: Cost-Sensitive Learning to Defer to Multiple Experts with Workload Constraints (mar 2024), <http://arxiv.org/abs/2403.06906>
2. Charusaie, M.A., Mozannar, H., Sontag, D., Samadi, S.: Sample Efficient Learning of Predictors that Complement Humans. In: Proceedings of the 39th International Conference on Machine Learning. pp. 2972–3005 (jul 2022), <http://arxiv.org/abs/2207.09584>
3. Cordelia, L.P., Stefano, C.D., Tortorella, F., Vento, M.: A Method for Improving Classification Reliability of Multilayer Perceptrons. *IEEE Trans. Neural Networks* **6**(5), 1140–1147 (1995). <https://doi.org/10.1109/72.410358>
4. De, A., Koley, P., Ganguly, N., Gomez-Rodriguez, M.: Regression under human assistance. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. pp. 2611–2620 (sep 2020) <https://doi.org/10.1609/aaai.v34i03.5645>, <http://arxiv.org/abs/1909.02963>
5. De, A., Okati, N., Zarezade, A., Gomez-Rodriguez, M.: Classification Under Human Assistance. In: The 35th AAAI Conference on Artificial Intelligence (AAAI-21) (2021) <https://doi.org/10.1609/aaai.v35i7.16738>, <http://arxiv.org/abs/2006.11845>
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. 33rd International Conference on Machine Learning, *ICML 2016* **3**, 1651–1660 (2016)
7. Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., Davis, J.: Machine learning with a reject option: a survey. *Machine Learning* **113**(5), 3073–3110 (jul 2024) <https://doi.org/10.1007/s10994-024-06534-x>, <http://arxiv.org/abs/2107.11277>

8. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Science Department, University of Toronto, Tech. pp. 1–60 (2009) <https://doi.org/10.1.1.222.9220>
9. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* **2017-Decem**, 6403–6414 (2017)
10. Leitão, D., Saleiro, P., Figueiredo, M.A.T., Bizarro, P.: Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. In: *International Conference on Machine Learning, Workshop on Human-Machine Collaboration and Teaming* (jun 2022), <http://arxiv.org/abs/2206.13202>
11. Madras, D., Pitassi, T., Zemel, R.: Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In: *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*. pp. 6150–6160 (nov 2018), <http://arxiv.org/abs/1711.06664>
12. Mao, A., Mohri, M., Zhong, Y.: Principled Approaches for Learning to Defer with Multiple Experts. In: *International Symposium on Artificial Intelligence and Mathematics (ISAIM 2024)* (2023), <http://arxiv.org/abs/2310.14774>
13. Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., Sontag, D.: Who Should Predict? Exact Algorithms For Learning to Defer to Humans. *Proceedings of Machine Learning Research* **206**, 10520–10545 (2023)
14. Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: *37th International Conference on Machine Learning, ICML 2020*. vol. PartF16814, pp. 7033–7044 (2020), <http://arxiv.org/abs/2006.01862>
15. Peterson, J., Battleday, R., Griffiths, T., Russakovsky, O.: Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision* **2019-Octob**, 9616–9625 (2019). <https://doi.org/10.1109/ICCV.2019.00971>
16. Verma, R., Barrejón, D., Nalisnick, E.: Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, PMLR 206. pp. 11415–11434 (2023), <http://arxiv.org/abs/2210.16955>
17. Wilder, B., Horvitz, E., Kamar, E.: Learning to Complement Humans. In: *IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. pp. 1526–1533 (may 2020), <http://arxiv.org/abs/2005.00582>



Explore Statistical Properties of Undirected Unweighted Networks from Ensemble Models

Xunda Zhao¹, Xing Wu², and Jianjia Wang³(✉) 

¹ School of Computer Engineering and Science, Shanghai University, Shanghai, China
zhaoxunda@shu.edu.cn

² Engineering and Science Shanghai Institute for Advanced Communication
and Data Science, School of Computer, Shanghai University, Shanghai,
People's Republic of China
xingwu@shu.edu.cn

³ School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University,
Suzhou, China
Jianjia.Wang@xjtlu.edu.cn

Abstract. Complex network theory has been widely demonstrated as a powerful tool in modeling and characterizing various complex systems. In the past, complex network theory has focused on the behaviors as well as the characteristics of the network nodes and edges. However, with the continuous evolution of society, traditional graph theory faces challenges due to the emergence of extremely large network structures. Recently, complex network method based statistics has attracted much attention. The new approach effectively manages very large networks and uncovers their intrinsic properties. In this paper, we present a complex network analysis model for undirected, unweighted networks based on a statistical analysis approach. This model is inspired by the ensemble model in thermostistical physics. Based on the established mathematical model, we derive physical measures that reflect the intrinsic properties of the network, including Entropy, Free Energy, Temperature, and so on. In the experimental part, we first explored the mathematical characterization of these metrics. Then, we observed the performance of various network categories under the same metric. Finally, we applied these measures to the field of graph classification. Extensive experiments demonstrate the effectiveness and superiority of the proposed method.

Keywords: statistical mechanics · complex network · entropy · ensemble

1 Introduction

The complex network theory provides a useful thinking method for analyzing complexity science and complex system [1]. They possess the capacity to depict,

elucidate, and model intricate real-world systems from both theoretical and practical perspectives [2]. For instance, a cell can be conceptualized as a complex network of chemicals linked by chemical reactions, the Internet is a complex network of routers and computers interconnected through diverse physical or wireless links [3] and a community as a complex network of interpersonal relationships woven by individuals who coexist. Indeed, myriad scientific, technological, social, economic, and biological systems can be elegantly and efficiently represented as networks: individual elements assume the role of nodes, and connections between these elements manifest as edges within the network, the edge weights can signify the strength of these connections, and unidirectional associations find expression in directed networks [4].

Utilizing classical graph-theoretic approaches, researchers have discovered numerous properties of complex systems, the best known of which are graph models, such as the small-world network models and the scale-free network models. The salient characteristic of scale-free networks is their adherence to a power law distribution in the degree distribution. This shows that most nodes exhibit minimal connections, while a scant few nodes facilitate the majority of interconnections [5]. The key property of small-world networks lies in the gradual expansion of the shortest path between two vertices with the scale of networks, often exhibiting logarithmic growth [6] [7]. This equates to a relatively short average distance between nodes in the network [8]. Exemplifying this concept, social networks stand as a quintessential instance of a small-world network. Real-world networks frequently showcase a community structure as well, characterized by subsets of vertices densely interconnected within, yet sparsely linked between [8] [9].

Despite relying on conventional methods, numerous complex network models have been constructed to describe complex systems. The complexity of traditional analysis methods inevitably increases with the network size, necessitating the development of more efficient analytical tools. Statistically based methods for analyzing complex networks were first proposed by Newman [10]. He applied the ensemble models from thermostistical physics to predict network properties and derive partition functions for random graphs and generalized random graphs. This set the stage for the subsequent development of statistically-based complex network models. Grounded in robust mathematical theory, statistical mechanics boasts significant utility in the examination of physical systems like fluids and solids. Its applicability to network research is equally justifiable [10]. Complex networks are usually evolving dynamically and typically encompass a substantial number of nodes and interconnected edges. The analogy with thermodynamic systems provides a rational basis for employing statistical mechanics methods in network science [11].

There are already numerous statistical models that yield physical measures characterizing the network. Richard C. et al. [12] applied the Maxwell-Boltzmann partition function to calculate network Entropy and model the Energy of the network in terms of thermodynamic equilibrium. For random graphs, small-world networks, and scale-free networks, they propose three physical measures

to characterize the intrinsic properties of networks: average network Energy, Free Energy, and network Temperature. However, the calculation of thermodynamic Entropy requires the network system to be in thermodynamic equilibrium. Hao-ran Zhu et al. [13] proposed the concepts of inverse Temperature and total Energy for the network, focusing on the variation of network Entropy with respect to network degree. Xin Zhao et al. [14] focused on the total Energy of edges in the network, which is primarily used to extract the network’s backbone structure.

In this paper, we propose a feature extraction model for undirected unweighted networks, also based on the ensemble model in statistical physics. We maximize the Entropy with the assumption that the number of nodes and edges in the network is fixed, and then we obtain the degree distribution of the nodes according to Maxwell-Boltzmann partition function. By analogy with physical measures in statical mechanics, six metrics characterizing the intrinsic patterns of the network are obtained: Temperature, Entropy, Free Energy, Capacity, Gibbs Free Energy, and Gibbs Chemical Potential. For example, the network Temperature indicates the average node degree, the Free Energy reflects the network’s ability to establish new connections, and the Gibbs Chemical Potential can be used to control the number of network nodes at a given Temperature.

In experiments, we calculated the metrics for each network in the dataset and analyzed their distributions. By fitting these distributions, we found that the metrics follow similar distribution patterns in pairs. For example, both network Entropy and Gibbs Free Energy conform to a normal distribution. We then explored the differences between various network categories under the same metric. For example, social networks exhibit smaller network Entropy compared to biological networks. Furthermore, to verify whether these network metrics can represent the unique characteristics of a network, we visualized the network clustering results using these six metrics as embedding. The results showed that these network metrics effectively differentiate between network categories. Finally, we used the network metrics for the graph classification task and compared the results with those obtained from a GCN [14] model on a graph dataset with no node or edge features. We found that using the network metrics as embeddings for the classifier yielded performance that was comparable to or better than that of the two-layer GCN model. This demonstrates that these network metrics can efficiently extract the intrinsic features of the network.

2 Related Work

Our paper mentioned graph model and ensemble model. In what follows, we provide a brief overview on related work in both models.

Classical Graph Model A regular graph is the simplest network structure, which has each node connected to all other nodes, which makes the degree of each node the same. Such a graph is also called a complete graph.

Random graphs have fixed values for certain attributes, while the structure of such networks is random in other respects. The simplest case involves a fixed

number of nodes and edges, but the positions of the edges connecting these nodes are randomized. The most famous random graph is the ER graph [15]. ER graph is constructed by randomly selecting E pairs of vertices from N vertices, each pair of vertices being selected with equal probability. Random graphs can be used to study the properties of the probability space associated with a graph with N nodes as N tends to infinity.

The property of small-world networks is that the shortest path between two vertices gradually expands as the network size increases, usually showing logarithmic growth. The two most common small-world network construction models are the WS model [16] and the NW model [17]. Small-world network models are widely used in social network analysis.

One other characteristic of real-world networks that have not been observed in previously mentioned model are growth. Scale-free networks are open to growth, as the most real networks. As the time passes, the number of nodes will be greater which is the same for World Wide Web that grows exponentially by addition of new web pages [18]. As a result, the node degrees of scale-free networks often follow a power law distribution.

Ensemble Model The concept of ensembles in statistical physics, which entails an assemblage of numerous potential states of a system under specific conditions [19]. In this paper, we deal with the canonical ensemble. The canonical ensemble represents the set of all possible states of a system in thermal equilibrium when in contact with a thermostatic thermal reservoir. The possible microstates of the system can have different energies because the system can exchange energy with the heat reservoir. In analogy with canonical ensemble, each node and edge in a complex network has different properties and the entire network is in contact with the real world at all times thus reaching thermal equilibrium.

3 Description Of Networks

Let $\mathcal{G}(V, \mathcal{E})$ be an unweighted and undirected network with a set of nodes V and a set of edges $\mathcal{E} \subseteq |V| \times |V|$. $N = |V|$ and $L = |\mathcal{E}|$ represents the total number of nodes on graph $\mathcal{G}(V, \mathcal{E})$. The adjacency matrix \mathbf{A} is defined as

$$\mathbf{A}_{uv} = \begin{cases} 1 & \text{if } (u, v) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where (u, v) is a pair of nodes forming an edge in the network.

The corresponding degree matrix \mathbf{D} is diagonal, where the elements are the degrees of the nodes,

$$\mathbf{D}(u, u) = k_u = \sum_{v \in V} \mathbf{A}_{uv} \quad (2)$$

where the degree of node u is k_u .

The average deegree \bar{k} , also known as the edge density, is

$$\bar{k} = \frac{1}{N} \sum_{v \in V} \mathbf{D}(u, u) = \frac{2L}{N} \tag{3}$$

For a network, the straightforward statistical property is the degree distribution, where n_k refers to the number of nodes which have certain value of degree k in the network. Thus,

$$\sum_{k=0}^{|V|} n_k = N \tag{4}$$

The preliminary Entropy relates to the number of ways for choosing n_k nodes among the total number of nodes N in the network. This is given by the combinatorial formula in terms of the factorials

$$\Omega = \frac{N!}{\prod_{k=0}^{|V|} n_k!} \tag{5}$$

Then, the Entropy is the lograithm of the expression Ω in (5) can be simplified by using Stirlings approximation $\log n! \approx n \log(n) - n$ and as a result

$$S = \log \Omega = -N \sum_{k=0}^{|V|} P_k \log P_k \tag{6}$$

where P_k is the nodal degree distribution of nodes as

$$P_k = \frac{n_k}{N} \tag{7}$$

Thus, the Entropy in a network is related to the degree distribution.

4 Method

4.1 Canonical Ensemble

The definition of canonical ensemble in our study refers to a very large group of networks with the identical number of nodes N and edges L . This means the network group follows an ergodic process, which contains all kinds of statistical structural propertites with the same number of nodes and edges. Our work is to study one of the graphs in this ensemble, particularly focus on one general case of a network with nodes N and edges L .

In a network, each node has the number of edges to connect that is a degree k_i , where $k_i \in 0, 1, 2, \dots, N$. The number of nodes with the certain degree k in the network is n_k , where $n_k \in 0, 1, 2, \dots, N$. Then, we want to maximise the Entropy in (6) with respect to the probability P_k , subject to two constraints

where the overall number of nodes and edges in a network is equal to N and L , respectively.

$$\begin{aligned}
 \mathbf{max} \log\Omega &= - \sum_{k=0}^{|V|} P_k \log P_k \\
 \mathbf{s.t.} \quad &\sum_{k=0}^{|V|} P_k = 1 \\
 &\sum_{k=0}^{|V|} k P_k = \bar{k} = \frac{2L}{N}
 \end{aligned} \tag{8}$$

A straightforward method is to apply Lagrange multipliers to find the optimum solution. This introduces two parameters α and β as

$$F = - \sum_{k=0}^{|V|} P_k \log P_k + \alpha \left(1 - \sum_{k=0}^{|V|} P_k \right) + \beta \left(\bar{k} - \sum_{k=0}^{|V|} k P_k \right) \tag{9}$$

When the partial differential of (9) with the respect of P_k is equal to zero, as

$$\frac{\partial F}{\partial P_k} = -\log P_k - 1 - \alpha - \beta k = 0 \tag{10}$$

We derive the degree probability by maximising the Entropy as

$$P_k = e^{-(1+\alpha)} e^{-\beta k} = \frac{1}{Z} e^{-\beta k} \tag{11}$$

where Z is the named as partition function following the first constraint that the overall probability is unit.

$$Z = e^{1+\alpha} = \sum_{k=0}^{|V|} e^{-\beta k} = \frac{1 - e^{-\beta N}}{1 - e^{-\beta}} \tag{12}$$

When the number of nodes in a network is large, the corresponding partition function in (12) can be approximated as

$$Z = \lim_{N \rightarrow \infty} \sum_{k=0}^N e^{-\beta k} = \frac{1}{1 - e^{-\beta}} \tag{13}$$

The second constraint of network edge proposes the relation of partition function and the average degree.

$$\bar{k} = \frac{2L}{N} = - \frac{1}{Z} \frac{\partial Z}{\partial \beta} = - \frac{\partial \log Z}{\partial \beta} = \frac{1}{e^\beta - 1} \tag{14}$$

According to (13), we derive the expression of parameter β as

$$\beta = \log \left(1 + \frac{1}{\bar{k}} \right) = \log \left(1 + \frac{N}{2L} \right) \quad (15)$$

The parameter β is named as inverse Temperature, and the corresponding Temperature T is given as

$$T = \frac{1}{\beta} = \left[\log \left(1 + \frac{N}{2L} \right) \right]^{-1} \propto \bar{k} \quad (16)$$

So the Temperature T is proportional to the average degree (edge density) in the network.

When combining (11), (13) and (15), we can derive the probability of a node with a certain degree as the function of degree distribution

$$P(k) = \frac{e^{-\beta k}}{Z} = (1 - e^{-\beta})e^{-\beta k} \quad (17)$$

which exponentially decays as the value of connected edges increasing. According to (15) and (17), we can also get the expression of degree probability in terms of the edge density as

$$P(k, \bar{k}) = \left(\frac{1}{\bar{k} + 1} \right) \left(\frac{\bar{k}}{\bar{k} + 1} \right)^k = \theta(1 - \theta)^k \quad (18)$$

where $\theta = \frac{1}{\bar{k} + 1}$.

This shows that the degree distribution is also a geometric distribution related to the average degree.

4.2 Limit Case of Canonical Ensemble

As the number of nodes N tends to infinity, with the partition function of (12) in hand, the preliminary entropy in the canonical ensemble case in (7) can be written as

$$\begin{aligned} S &= - \sum_{k=0}^{|V|} P_k \log P_k = - \sum_{k=0}^{|V|} \frac{e^{-\beta k}}{Z} (-\beta k - \log Z) \\ &= \beta(\bar{k} - F) \end{aligned} \quad (19)$$

where F is named as the Helmholtz Free Energy in statistical mechanics as

$$F = -\frac{1}{\beta} \log Z = -T \log Z \quad (20)$$

It reflects the ability of the network to make new connections.

To simplify, the parameter μ in the grand canonical ensemble can be computed from Helmholtz free energy as

$$\mu(T) = \left(\frac{\partial F}{\partial N} \right)_T = \frac{1}{1 - e^{N/T}} \quad (21)$$

Therefore, the parameter μ is named as Chemical Potential that is used to control the number of nodes in the network corresponding to the Temperature.

From (16), we can also derive the Capacity C which reflects how the average degree in a single node changes with the Temperature as

$$C = \left(\frac{\partial \bar{k}}{\partial T} \right) = \frac{e^{1/T}}{T^2(e^{1/T} - 1)^2} \quad (22)$$

This show the ability of a node to attract new connections is not always increasing with the degree. It can reach to a platform while the connection is saturation in a network.

So far, we only consider a single node in a network with the statistical mechanics framework. For a network with N indistinguishable nodes, the partition function in (12) can be written as Gibbs partition function

$$\mathcal{Z} = \frac{Z^N}{N!} \quad (23)$$

Then, the Helmholtz Free Energy can be calculated when combine with the partition function in (13)

$$\mathcal{F} = -T \log(\mathcal{Z}) = -TN \log Z + T \log(N!) \quad (24)$$

The corresponding Chemical Potential in (21) is

$$\mu = \left(\frac{\partial \mathcal{F}}{\partial N} \right)_T = T \log \left(\frac{N}{Z} \right) = T \log \left(\frac{N}{1 + \bar{k}} \right) \quad (25)$$

We can observe the critical point happens when $\mu = 0$. This is consistent to be trivial solution that number of edges in a network reach to a half of potential connections in nodes.

$$\bar{k} = N - 1, \quad \text{when } \mu = 0 \quad (26)$$

that is

$$L = \frac{1}{2}N(N - 1) \quad (27)$$

Therefore, the parameter μ can be used to reflect the edge density changes in the network connection.

5 Experiments

5.1 Datasets

We use five well-known datasets consisting of two complex network repository (KONECT [20] and REPOSITORY [21]) and three graph classification

Table 1. Statistics of the experimental benchmark datasets. #G denotes the numbers of graphs. #Ctg denotes the numbers of categories #Cls denotes the number of class labels. Avg#N denotes the average number of nodes per graph. Avg#E denotes the average number of neighbors per node. d is the dimension of feature vectors.

Dataset	#G	#Ctg	#Cls	Avg#N	Avg#E	d
KONECT	1,326	24	-	-	-	-
REPOSITORY	5,661	33	-	-	-	-
COLLAB	5,000	1	3	74.5	65.9	-
IMDB-B	1,000	1	2	19.8	9.8	-
DD	1,178	1	2	284.3	5.0	82

datasets(COLLAB, IMDB-B, and DD). The COOLAB, IMDB-B datasets do not have available node features; thus we use node degrees as features. Dataset statistics are summarized in Table. 1.

KONECT: The Koblenz Network Collection, one of the most important dataset in the field of network science, contains over 1,000 network datasets. KONECT contains networks of all sizes, from samll calssical datasets from the social science, such as Kenneth Tead’s Highland Tribes network with 16 vertices and 58 edges, to the Twitter social network with 52 million nodes and 1.9 billion edges.

REPOSITORY: The largest network repository with thousands of donations in 30+ domains from biological to social network data.

Social networks datasets: COLLAB is a scientific dataset, where each graph represents a collaboration network of a corresponding researcher with other researchers from each of 3 physical physics fields; each graph is labeled to a physics field that the researcher belongs to. IMDB-B is movie collaboration datasets, where each graph is derived from actor/actress and genre information of different movies on IMDB.

Bioinformatics datasets: DD is a collection of 1,178 protein network structures with 82 discrete node labels, where each graph is classified into enzyme or non-enzyme class.

5.2 Experimental Results

Experiment 1 In this experiment, we explored the mathematical distributional properties of the proposed six metrics, Entropy, Temperature, Capacity, Free Energy, Gibbs Chemical Potential, and Gibbs Free Energy. First, we computed six metrics on KONECT and REPOSITORY, and we constructed statistical histograms of these results. Fig. 1 portrays the metrics distribution of the database KONECT. Note that we take logarithms with base ten for Gibbs Free Energy. Obviously, Entropy and Gibbs Free Energy seem to obey a similar mathematical distribution, and the same phenomenon occurs between Temperature and Gibbs Chemical Potential, Capacity, and Free Energy. Fig. 2 shows the results calculated on the REPOSITORY and the distribution pattern between the measures

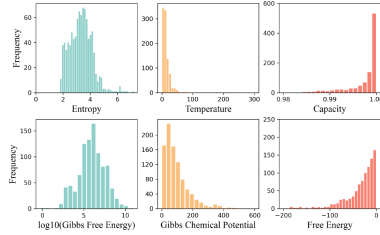


Fig. 1. Histogram of the distribution of the six metrics on the KONECT dataset. We can see that each column of metrics follows the same distribution. Note: we take logarithms with base ten for Gibbs Free Energy.

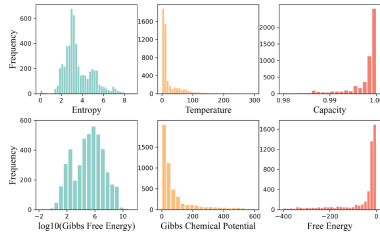


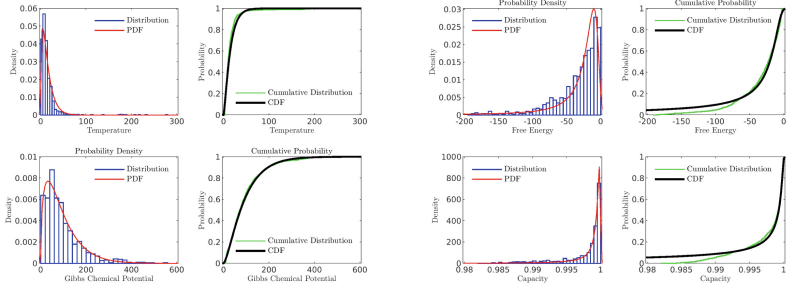
Fig. 2. Histogram of the distribution of the six metrics on the REPOSITORY dataset. Similarly, in this dataset, we can observe the same conclusion as mentioned in Fig. 1

is the same as that shown in Fig. 1. And the distribution of the same measure is also the same between the two datasets.

Then, we fit distributions to these metrics using KONECT as an illustration. The Results indicate that Entropy and Gibbs Free Energy follow a normal distribution, Temperature, and Gibbs Chemical Potential obeys the Gamma distribution, and Capacity and Free Energy follow the stable distribution. Fig. 3 exhibits the results of the probability density fitting curve and the cumulative probability distribution comparison.

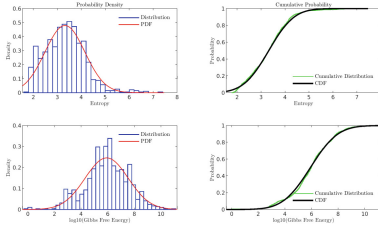
Experiment 2 Real-world networks can be loosely classified into four categories: social networks, information networks, technological networks, and biological networks. Notably, social and biological networks have garnered significant research attention [22]. In this experiment, we use Entropy and Gibbs Chemical Potential as measures to explore the differences between social networks and biological networks. In Fig. 4a, it becomes evident that biological networks exhibit lower Entropy in comparison to social networks. We believe that the reason for this difference is that the structure of biological networks is relatively regular. Such structure is determined by the genetic and metabolic patterns which all creatures have similarly. In contrast, social networks tend to evolve more randomly, rendering them inherently less organized.

The Gibbs Chemical Potential shows the tendency of the Free Energy of the network to change with the number of nodes when given a specific Tempera-



(a) Temperature and Gibbs Chemical Potential are approximately Gamma distribution.

(b) Capacity and Free Energy are approximately Stable distribution.



(c) Entropy and Gibbs Free Energy are approximately Normal distribution.

Fig. 3. Probability density function fitting and cumulative metric distribution for six metrics on KONECT. In Fig. 3b, the value of the fitted cumulative probability distribution is small compared to the actual cumulative probability distribution when Free Energy is less than -60 . But the number of networks is few in this interval, we considered the fitting results to be valid. The same is true for capacity.

ture. Fig. 4b displays that the Gibbs Chemical Potentials of biological networks generally exceed social networks. This discrepancy could be attributed to the small-world property that is common in social networks. The new nodes added into the community have little impact on the integral network structure. As a result, in social networks, the variation of Free Energy caused by the change in the number of nodes is not obvious.

Experiment 3 In this experiment, we explored the feasibility of the proposed metrics as network embeddings. We performed dimensionality reduction on the network features using the TSNE method on the REPOSITORY. For ease of presentation, we have listed only the top 18 categories with the highest number of datasets. In Fig. 5, It can be seen that the result of using the six metrics as network features to cluster is basically the same as the original network categories.

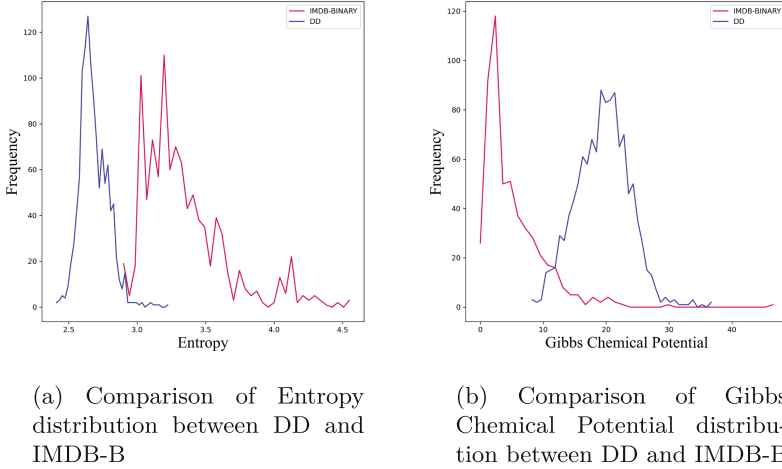


Fig. 4. Comparison of biological and social networks under Entropy and Gibbs Chemical Potentials on DD and IMDB-B, respectively. The results show that social networks have greater Entropy, while biological Networks have greater Gibbs Chemical Potential.

Table 2. Graph classification results(% accuracy). The best scores are in bold

Model	COLLABIMDB-BDD		
GCN-2	66.6%	62.03%	41.52%
Our*	67.69%	67.98%	77.39%
DGCNN [23]	68.34%	-	77.21%
GAP-Layer(Ncut) [24]	65.89%	68.80%	-
1-NMFPool [25]	65.0%	-	76.00%
CT-Layer [24]	69.87%	69.84%	-

* using proposed metrics as embeddings feeds to single-layer classifier.

Experiment 4 Inspired by the results of Fig. 5, in this experiment, we verified the feasibility of the proposed metric for supervised graph classification tasks. We used the proposed metrics as the graph embeddings feeding to single-layer classifier. As a contrast, we selected some graph convolution models to generate graph embeddings and then applied them to classifier of the single-layer structure. For two datasets, COLLAB and IMDB-B, there are no node features, so we use one-hot vectors of node degree as node features. The DD dataset uses one-hot vectors of node labels as node features. Table .2 presents the experimental results for the benchmark datasets. It can be seen that although our classification accuracy is not the highest in COLLAB and IMDB-B, it is sufficient to demonstrate our proposed metrics can be used as graph embeddings for downstream tasks.

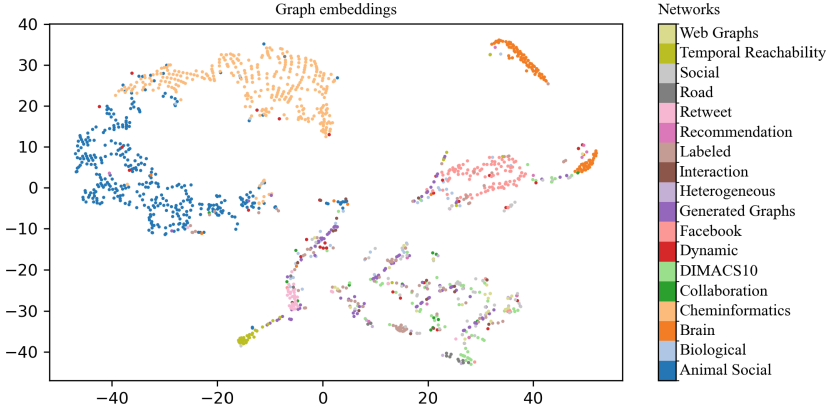


Fig. 5. A T-SNE visualization using proposed metrics as graph embeddings on REPOSITORY.

6 Conclusion

This paper introduces a model based on statistical mechanics, applied for unweighted and undirected networks, which yields six parameters characterizing the features of the network. Inspired by statistical mechanics canonical ensemble concepts, we employ the maximum Entropy method to study networks with fixed count of nodes and edges. Through the Lagrange multiplier technique, we proved that the network’s degree distribution adheres to a geometric distribution under maximal Entropy conditions. In line with thermodynamic principles, we introduce the notion of network Temperature and deduce the distribution function for the canonical ensemble.

In the experimental part, we designed four experiments. In the first experiment, we focused on the mathematical distribution of the proposed metrics and determined the statistical distribution followed by the metric through distribution fitting. In the second experiment, we explored the difference between social and biological networks in terms of Entropy and Gibbs Chemical Potential, and found that the Entropy of social networks is small, while the biological Entropy is on the large side. In the third experiment, we clustered the network data by TSNE, and the experimental results are similar to the original network categories, proving that the proposed metrics can be used as network embedding. In the fourth experiment, we take six metrics as graph embedding and use them for the graph classification task, and compared with the two-layer GCN model, directly using the proposed metrics for the classifier works better than the two-layer GCN model in the dataset without node features.

Although our proposed metrics can extract network macroscopic features, it is unable to take into account network microtopology and graph-level features when compared to the graph convolutional models. In future work, we will explore the combination of the proposed metrics and graph convolutional models

to improve the graph classification performance in graph datasets without node features.

Acknowledgements. This work is supported by the Natural Science Foundation of Jiangsu Higher Educational Institution of China(Grant No.24KJB510049), and Research Development Fund (RDF-23-01-044) at Xi'an Jiaotong-Liverpool University.

References

1. Cheng, Y., Sun, F., Zhang, Y., Tao, F.: Task allocation in manufacturing: A review. *J. Ind. Inf. Integr.* **15**, 207–218 (2019)
2. da Fontoura Costa, L.: Coincidence complex networks. *Journal of Physics: Complexity* **3**(1), 015012 (mar 2022)
3. Statistical mechanics of complex networks: Albert, R.k., Barabási, A.L.s. *Rev. Mod. Phys.* **74**, 47–97 (2002)
4. Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., Caldarelli, G.: The statistical physics of real-world networks. *Nature Reviews Physics* **1**(1), 58–71 (2019)
5. Broido, A.D., Clauset, A.: Scale-free networks are rare. *Nature communications* **10**(1), 1017 (2019)
6. Barrat, A., Weigt, M.: On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems* **13**, 547–560 (2000)
7. Amaral, L.A.N., Scala, A., Barthélemy, M., Stanley, H.E.: Classes of small-world networks. *Proc. Natl. Acad. Sci.* **97**(21), 11149–11152 (2000)
8. Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**(12), 7821–7826 (2002)
9. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113 (2004)
10. Park, J., Newman, M.E.: Statistical mechanics of networks. *Phys. Rev. E* **70**(6), 066117 (2004)
11. Pathria, R.K.: *Statistical mechanics*. Elsevier (2016)
12. Wang, J., Wilson, R.C., Hancock, E.R.: Network entropy analysis using the maxwell-boltzmann partition function. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 1321–1326. IEEE (2016)
13. Zhu, H., Wu, H., Wang, J., Hancock, E.R.: Weighted network analysis using the debye model. In: *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2020, Padua, Italy, January 21–22, 2021, Proceedings*. pp. 153–163. Springer (2021)
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations* (2017)
15. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *nature* **393**(6684), 440–442 (1998)
16. On the evolution of random graphs: Erdős, P., Rényi, A. *Publ. Math. Inst. Hungar. Acad. Sci* **5**, 17–61 (1960)
17. Newman, M.E., Watts, D.J.: Renormalization group analysis of the small-world network model. *Phys. Lett. A* **263**(4–6), 341–346 (1999)
18. Kaviani, S., Sohn, I.: Application of complex systems topologies in artificial neural networks optimization: An overview. *Expert Syst. Appl.* **180**, 115073 (2021)

19. Brede, M.: Networks-an introduction. mark ej newman.(2010, oxford university press.) isbn-978-0-19-920665-0. (2012)
20. Kunegis, J.: Konect: the koblenz network collection. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 1343–1350. WWW '13 Companion, Association for Computing Machinery, New York, NY, USA (2013)
21. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI (2015)
22. Newman, M.E.J.: The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003)
23. Zhang, M., Cui, Z., Neumann, M., Chen, Y.: An end-to-end deep learning architecture for graph classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
24. Arnaiz-Rodríguez, A., Begga, A., Escolano, F., Oliver, N.M.: Diffwire: Inductive graph rewiring via the lovász bound. In: Learning on Graphs Conference. pp. 15–1. PMLR (2022)
25. Bacciu, D., Di Sotto, L.: A non-negative factorization approach to node pooling in graph convolutional neural networks. In: AI IA 2019–Advances in Artificial Intelligence. pp. 294–306. Springer (2019)



Evaluation of Machine Learning Techniques for Classification of Surface Roughness of Machined Samples using Laser Speckle Imaging Technique

Shanta Hardas Patil^(✉) 

Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India
shantahardas@iitg.ac.in

Abstract. This study uses machine learning techniques to classify the surface roughness using the laser speckle images of the machined samples, an intriguing yet relatively less explored field of research in the realm of speckle metrology. The laser speckle imaging technique is sensitive to surface roughness paving the way for the classification of the machined specimen based on surface roughness using the distinct speckle pattern. The paper presents the analysis of the performance of the state-of-the-art machine learning techniques on the preliminary dataset of the speckle pattern of the machined sample. The gray level co-occurrence matrix is used for feature extraction. The model performance with various combinations of features is studied to distinguish the most descriptive feature for generalization. The assessment of the classifiers' performance aids in the generalization of the classification and prediction of the roughness classes in the range $R_a = 0.1 \mu\text{m} - 1.6 \mu\text{m}$ using the speckle images.

Keywords: Surface Roughness · Laser Speckle Imaging Technique · Classification · Gray Level Co-Occurance Matrix · k-Nearest Neighbour · Support Vector Machine · Decision Tree · Random Forest

1 Introduction

Surface Roughness, R_a is an important texture property that specifies the microstructural variation of the material surface. For industrial applications, it is pertinent to understand the micro-details of a surface to match the desired surface texture characteristics. The laser speckle imaging technique (LSI) [16, 17] is a non-destructive optical imaging modality for gathering multiparameter object information. *Speckle Metrology*, based on LSI techniques, is a new paradigm for numerous industrial applications including measurement of object movement, vibration modes, and surface roughness [14, 25]. The *speckles* represent the intensity distribution developed from the mutual interference of the wavefront scattering from the object when illuminated by a coherent light source. Since the

wavefront scattering takes place from the different depths of the object's surface, the statistical analysis of speckle patterns displays subtle variations related to the surface roughness. Therefore, the speckle imaging technique is relatively suitable for the classification and prediction of surface roughness measurement compared to the other non-destructive techniques.

The statistical methods for speckle image processing were proposed including Speckle Contrast(SC) [2, 13] which is frequently used for a small range of R_a measurements. Besides SC, bright to dark (B/D) pixel ratio [9, 20], speckle correlation [18], fractal dimensions [7], Hurst exponent [1] were proposed for R_a measurement. With the digital transition, the role of machine learning techniques for speckle image processing to convert the observations into the desired attributes is pivotal to accelerate the performance and alleviate the pitfall of the traditional image-processing algorithms.

The classification and prediction of surface roughness of the machined samples based on machine vision technique are reported in [4, 5, 15, 22, 24]. Nevertheless, machine vision is a non-destructive and undisturbing method of data acquisition, it is very sensitive to the lighting conditions. In dim or dark conditions, the quality of the image can degrade affecting the true results. In such cases, artificial lighting is an additional requisite. Moreover, for unstructured scenes, specimen identification may be difficult with the machine vision technique. In comparison, the laser speckle images can encode the microstructural details of the surface from the scattering of the light, making speckle patterns distinct for each surface that visually might appear alike. Therefore, the prediction and classification of surface roughness of the optically smooth surfaces is feasible with LSI. In addition, the laser speckle imaging technique is suitable for 2-dimensional and 3-dimensional surfaces where machine vision techniques might face several limitations.

In this paper, k-Nearest Neighbour, Support Vector Machine, Decision Tree, and Random Forests techniques have been used to illustrate the surface roughness-based classification of the machined samples using the speckle images. The samples are prepared using the grinding operation. The classification for the five classes of grinding samples is performed for $R_a = 0.1\mu m - 1.6\mu m$ with a brief discussion on the classification accuracy and the important features. Section 1 presents the introduction, methodology is described in Section 2; Section 3 covers the experiment details followed by results and discussion in Section 4; Section 5 discusses the potential scope for improvement, and Section 6 reports the conclusion and future work.

2 Methodology

2.1 Feature Extraction

The speckle is a 2-dimensional random intensity pattern, therefore they are often investigated by statistical methods. Gray-level co-occurrence matrix (GLCM) is a second-order statistical method for texture analysis [11]. The statistic of pixel intensity distributions in the direction θ and distance d between the pixels is

obtained using GLCM. The pixel pair in GLCM with certain mutual spatial relationships provides more efficient representative information than a single pixel, thereby bringing GLCM to the front for the feature extraction of the images. GLCM calculates how often a pixel value i occurs either horizontally (0°), vertically (90°), or diagonally (45° and 135°) to the adjacent pixels of value j . Haralick et. al [11] defined fourteen features that provide spatial context information for texture classification. Connors and Harlow [6] reported energy, entropy, local homogeneity, and inertia as the few significant GLCM features, that are commonly used. The following procedure is followed for the feature extraction using GLCM:

- Decide the offset distance integer (d) and direction (θ).
- The formation of GLCM matrix for an image I of size $M \times N$; where each pixel-pair is separated by d and θ , the probability of having the pixel intensity i next to pixel intensity j is addressed. The GLCM is a square matrix of size N where N is the total number of grey levels in the image. Each matrix element is represented by $P(i, j|d, \theta)$.

After GLCM formulation, one can compute the statistical features given below:

– **Energy**

It is also known as the angular second moment (ASM). It is a measure of image uniformity.

$$ASM = \sum_{i=1}^N \sum_{j=1}^N [P_d(i, j)]^2. \quad (1)$$

– **Contrast**

Contrast is the difference between the highest and the lowest values of a continuous set of pixels. It is useful to measure the grey-level local intensity variations in the image.

$$Con = \sum_{i=1}^N \sum_{j=1}^N (i, j)^2 P_d(i, j). \quad (2)$$

– **Homogeneity**

It is measured to show the closeness of GLCM matrix elements to the GLCM diagonal.

$$Hom = \sum_{i=1}^N \sum_{j=1}^N \frac{1}{1 + (i - j)^2} P_d(i, j). \quad (3)$$

– **Correlation**

The joint probability occurrence of the specified pixel pairs is measured using the correlation parameter.

$$Corr = \sum_{i=1}^N \sum_{j=1}^N P(i, j) \frac{(i - \mu_x)(j - \mu_y)}{\sigma_x \sigma_y}. \quad (4)$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviation of rows and columns, respectively.

– **Entropy**

It is a measure to describe the disorder or the complexity of the image.

$$Ent = - \sum_{i=1}^N \sum_{j=1}^N P_d(i, j) \log_2 P_d(i, j). \quad (5)$$

– **Disssimilarity**

Dissimilarity represents the linear local variation in the image intensity.

$$Diss = \sum_{i=1}^N \sum_{i=1}^N |i - j| P_d(i, j). \quad (6)$$

where μ is a mean of $P_d(i, j)$.

2.2 Feature Selection

The appropriate selection of features is important for effective classification. For the classification of surface roughness based on the respective speckle images of the samples, the following aspects are considered for the feature selection:

- To reduce the dimensionality of feature space and avoid overfitting.
- To improve the accuracy of a classification algorithm.

According to the principle of laser speckle imaging, when the light is scattered from a random surface with higher roughness, the intensity of the speckle grain reduces, thus the resultant speckle pattern contains more dark speckles as compared to the smooth surface. Eventually, the number of similar neighbor pixels tends to increase which increases the homogeneity of the sample. Therefore, the correlation between the neighboring pixels increases. Moreover, the contrast of the speckle pattern decreases with an increase in surface roughness. Numerous studies have demonstrated the relation between the average surface roughness and the contrast of the speckle images based on the regression analysis. The correlation increases with an increase in the surface roughness. In addition to the aforementioned features, energy or angular second moment is weekly sensitive to the smooth surface, which varies between 0 – 1. Moreover, according to the general principle, the feature must be discriminative, reliable, and independent. Based on these considerations, the analysis is confined to the Haralick features viz., contrast, correlation, homogeneity, dissimilarity, entropy, and energy in the present study.

2.3 Classification Methods

k-Nearest Neighbour: k-Nearest Neighbor (kNN) is a supervised machine learning technique for classification due to its simple implementation and distinguished performance [23]. In the case of classification, kNN aims to find the

value of k on the observations in the training data which is closest to the observations in the testing datasets [26]. Being a user-defined constant, the effective classification of data explicitly depends on the appropriate selection of k -values. The Euclidean distance metric is most widely used in k NN defined below:

$$d(x, y) = \sqrt{\left(\sum_{i=1}^p (x_i - y_i)^2\right)}. \quad (7)$$

The other metrics include the Manhattan distance, the Chebyshev distance, the Cosine distance, the Jaccard distance, and the Hamming distance. The Minkowski distance is known as a generalized form used for the Chebyshev, the Manhattan, and the Euclidean distances calculation depending on the selection of p .

Support Vector Machine: Support vector Machine (SVM) was formally introduced by V. Vapnik in 1965. SVM is widely applied in pattern recognition, text and image classification, biological sequence analysis, natural language processing, etc [8]. The model uses a linear classifier with an optimal margin in the feature space. The learning strategy is to maximize the margin to formulate a convex quadratic programming problem. The input vector X is mapped into a high-dimensional decision feature space S non-linearly by choosing a priori. The decision surface is known as hyperplane H , defined as the maximum margin of separation between two classes [12].

$$H = \{\mathbf{x} | \mathbf{w}^T \mathbf{x} + b = 0\}, \quad (8)$$

where \mathbf{w} is a weight vector normal to the plane and b is a bias. The SVM can also be used on a dataset that might have required a non-linear decision plane. This is achieved using various kernel functions which can map the non-linearly separable input space into a linearly separable space. The kernel functions are described below:

- **Linear Kernel:** It assumes the linear relationship between the data points and is suitable for the dataset following a linear trend.

$$K(x, x_i) = [(x^T \cdot x_i)], \quad (9)$$

- **Polynomial Kernel:** Polynomial kernel is used for learning the non-linearity in the data points. It is expressed as follows:

$$K(x, x_i) = [x^T \cdot x_i + c]^d, \quad (10)$$

x and x_i are the vectors of size n in the input space. c is the constant used to trade off between the higher-order and lower-order terms in the polynomials. The degree of polynomial is represented by d .

- **Sigmoid Kernel:** The Sigmoid kernel provides the basis for similarity measure based on the hyperbolic tangent function.

$$K(x, x_i) = \tanh(\alpha(x \cdot x_i) + \beta), \quad (11)$$

The hyperparameters α and β control the shape of the kernel.

- **Gaussian Radial Basis Function Kernel:** This function uses the Euclidean distance relationship between the data points. The parameter g represents the width of the Gaussian function which aids the adoption of the high dimensional datasets of varying scale.

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{g^2}\right). \quad (12)$$

Decision Tree: The Decision Tree(DT) is a non-parametric supervised learning method that simplifies the relationships between feature input and the target variables by partitioning the original features into potential subgroups. It is often used to deal with large and complex datasets for classification, prediction, and data manipulation tasks. The decision tree methods can be used to select the most relevant features that should be used to form decision tree models. The model performance is sensitive to the splitting, stopping, and pruning[21].

Random Forest: Random Forest (RF) technique treats the overfitting problem of decision tree algorithm on training dataset through ensemble learning by constructing the multiple decision tree during the training phase. The subset of the feature is selected in a random split while the decision tree considers all the possible feature splits.

3 Experiment

3.1 Description of LSI Setup

The schematic of the experimental assembly is shown in Fig. 1. The sample was illuminated using a laser source at an illumination angle, $\theta_i = 45^\circ$ forming a speckle pattern due to wavefront interference at the image plane. The interferences of the wavefront captured by the CMOS sensor are recorded using the image acquisition software as a speckle pattern. The distance between a laser source and an object to the image sensor was kept unchanged during the experiment. A He-Ne laser source of wavelength $\lambda = 632.5 \text{ nm}$, 15 mW and a CMOS camera with a maximum resolution of 2048×1536 and pixel size $3.45 \text{ } \mu\text{m} \times 3.45 \text{ } \mu\text{m}$ were used for the illumination and recording of a speckle pattern, respectively. Fig. 2 shows the recorded speckle images for each class of R_a .

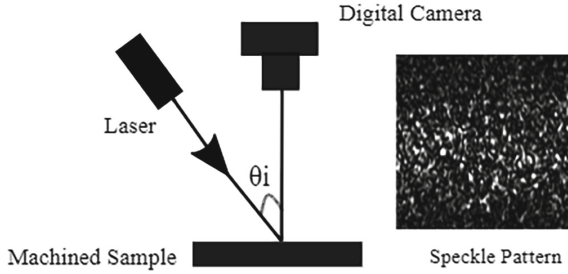


Fig. 1. Schematic of the experimental setup for the acquisition of speckle images of the machined samples.

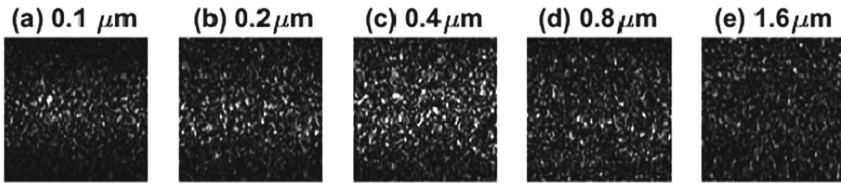


Fig. 2. Speckle images of the grinding sample classes with surface roughness R_a (a) 0.1 μm (b) 0.2 μm (c) 0.4 μm (d) 0.8 μm (e) 1.6 μm .

3.2 Data Collection and Data Description

The speckle images of the grinding specimen were recorded by scanning at different sampling positions of the specimen. The preliminary dataset consists of a total of 200 speckle patterns of five classes of surface roughness as tabulated in Table 1. The speckle of the respective class of roughness is distinct in terms of the speckle intensity and speckle size. The speckle is a random intensity pattern representing the structural irregularities of the corresponding point of illumination on the sample plane, it is noteworthy that the speckle pattern of the two closely sampled positions of the same class of roughness are not alike. For the classification analysis, the raw speckle image data is used for each classifier.

Table 1. The description of the sample set.

R_a (μm)	0.1	0.2	0.4	0.8	1.6
label	0	1	2	3	4

4 Results and Discussion

This section describes the result and analysis of experiments conducted for the classification of surface roughness of the grinding specimen based on the speckle

Table 2. The impact of feature selection on the accuracy. FS-Feature Subset; FS-I: contrast, correlation, homogeneity, dissimilarity, energy, entropy; FS-II: correlation, homogeneity, dissimilarity, energy, entropy; FS-III- dissimilarity, energy, entropy, homogeneity, FS-IV: correlation and dissimilarity, FS-V: correlation, dissimilarity, entropy; FS-VI: homogeneity, dissimilarity, energy. The cross-validation is obtained for $k = 10$ -fold. The accuracy is expressed in %.

FS-I		FS-II	
K-value	Cross -validation Test Accuracy	Cross-validation Test Accuracy	
3	72.50	75	99.00
5	75.50	78.33	98.5
7	72.99	76.66	97.00
9	75.50	80.00	96.00
11	74.5	78.33	93.99
13	75.50	75.50	93.49
15	74.49	73.33	92.5
17	73.00	71.66	95.5
19	76.49	68.33	92.99
FS-III		FS-IV	
K-value	Cross -validation Test Accuracy	Cross -validation Test Accuracy	
3	95.00	100	99.00
5	92.00	95.00	98.00
7	89.49	93.33	96.50
9	87.5	83.33	95.50
11	84.5	83.33	93.49
13	83.00	83.33	92.99
15	82.00	83.33	91.49
17	80.99	76.99	90.99
19	81.50	76.77	96.00
FS-V		FS-VI	
K-value	Cross -validation Test Accuracy	Cross -validation Test Accuracy	
3	99.00	100	89.50
5	98.00	98.33	85
7	97.00	98.33	82.5
9	96.00	96.33	80.5
11	93.99	95.00	76.5
13	93.49	95.0	78.49
15	92.50	96.66	78.00
17	92.50	95.00	75.50
19	96.00	95.00	78.00

images. The kNN algorithm is often sensitive to the selection of the k value [10]. Liu et al. suggested that a fixed k value may not be suitable for many test data points in a given training dataset[19]. Nevertheless, kNN is a simple algorithm, the high dimensional feature space hinders the performance of kNN[3]. In the present case, the effect of variable k value for different GLCM feature subsets (FS) which significantly affects the model's predictions is presented. The cross-validation for $k = 10$ is analyzed for the train-test split of 70:30 of the dataset. The importance of each feature is determined by examining how the performance is affected with or without having the specific feature(s). If the removal of feature degrades the classifier performance, the feature is considered to be important. The key contributing features are thus selected for the final feature subset. As described in Table.2, at first, the FS-I consisting of all six features resulted in 80% testing, and 75.50% cross-validation (CV) accuracy which is the highest for $k = 9$. In the FS-II configuration, the removal of contrast shows a remarkable increase in testing and cross-validation accuracy. The optimal value of k is 3 for which the CV = 99% and the test accuracy obtained is 100%. The accuracy further reduces with an increase in k. In the FS=III the CV accuracy is 95.50% while testing accuracy is 100% for $k= 3$. Due to the high difference in the testing and CV accuracy, FS-III seems unpredictable for generalization. The performance of kNN with a combination of correlation and dissimilarity (FS-V) shows a consistent decrease in CV with a marginal difference with the testing accuracy. To investigate further, the inclusion of entropy over FS-V has moderately increased the accuracy beyond $k= 7$. Furthermore, homogeneity, energy, and dissimilarity in FS-VI showed the cross-validation accuracy higher than the testing accuracy which might lead to overfitting with the unseen data. the kNN suffers from the curse of dimensionality with a large number of features as apparent in the case of the present dataset as well. For the generalization, dissimilarity, correlation, and entropy are promising descriptive features as compared to the other GLCM features for the kNN model. Fig.3 (A),(B),(C) illustrates the plot of accuracy (CV and Test) versus k for FS-I, IV, and, V. The plot in Fig 3 (D) represents accuracy (CV and Test) versus K for energy, entropy, and homogeneity which indicates the overfitting of the data. The FS-V seems promising for model building using kNN for the classification of surface roughness.

The performance of SVM with linear and the rbf kernel is presented in the Table. 3. The exhaustive search method is applied for the selection of optimal hyperparameters. The linear kernel with FS-II having correlation, dissimilarity, energy, homogeneity, and entropy features resulted in 100% accuracy, precision, recall, and F1-score for $C = 10$ whereas, the accuracy obtained with the rbf kernel is 98.33% for $C= 10$ and $\gamma = 1$ for the similar FS. The FS-I, FS-III, FS-IV, and FS-V under-perform indicating the incompetency for prediction on unseen data. The statistics presented in Table. 3 infer that the performance of FS-II with the linear kernel is reliable for the generalization as compared to the remaining combinations.

Fig. 4 shows the plot for the testing and training accuracy as a function of the maximum depth of the decision tree. The plot indicates the model performs well

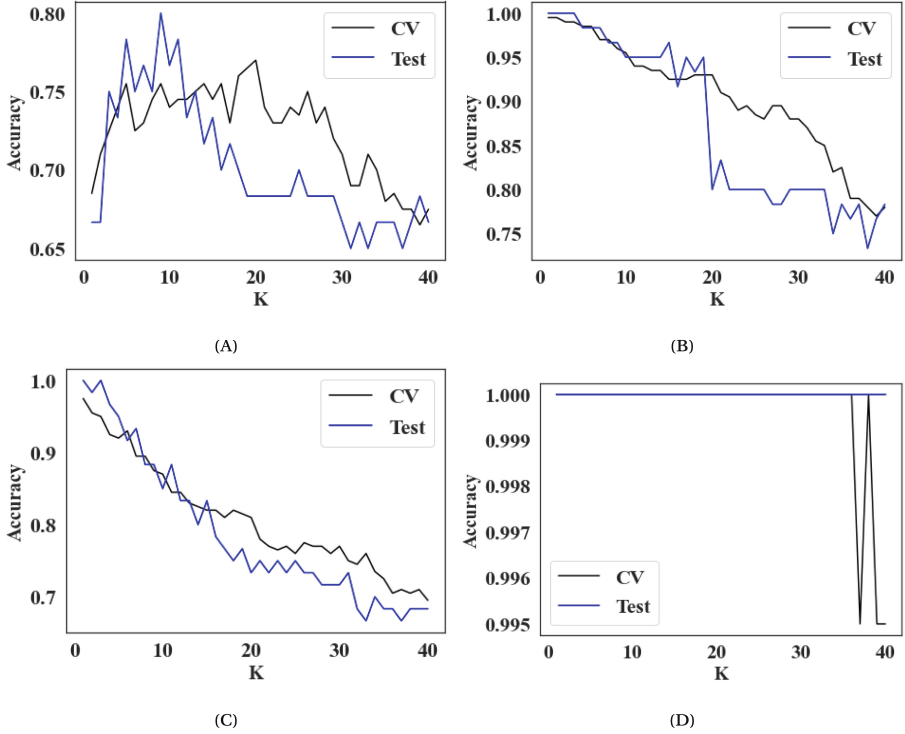


Fig. 3. The plot of cross-validation and test accuracy as a function of nearest neighbour, k for kNN model with (A) FS-I (B) FS-IV, (C) FS-V, and (D) The combination FS with energy, homogeneity, and entropy.

on the training data but is relatively poor on unseen instances which indicates the overfitting due to the small data size. The problem of overfitting can be countered with pruning implementation. From the plot, it is observed that the minimum level of depth required is 4 for which the training and validation accuracy are approximately similar. The performance of the random forest is analyzed based on the train-test split of the dataset in the ratio of 80:20, 70:30, and 60:40 using the cross-validation for k -fold= 5 and k -fold= 10 as tabulated in Table.4. The 10-fold CV accuracy obtained for the train-test split of 70:30 is 96.42%.

The Nested Cross-Validation(N-CV) procedure is often justified by providing a more reliable means of model selection for a given dataset. In the N-CV, the hyperparameter selection is performed in the inner cross-validation, whereas, the outer cross-validation provides an unbiased estimate of the expected accuracy of the algorithm. The statistics in Table.2 and Table.3 show the generalization performance obtained with CV accuracy is overly optimistic for certain FS. Therefore, N-CV is performed for those feature subsets. Table.5 presents the N-CV score for the kNN and SVM model. The SVM(linear) model with FS-II provides best performance for a given dataset with hyperparameters, $C=10$.

Table 3. The comparative analysis of SVM with the kernel function: linear and rbf for various Feature Subset (FS). All values are represented in %. FS-I: contrast, correlation, dissimilarity, energy, entropy, homogeneity; FS-II: correlation, dissimilarity, energy, entropy, homogeneity; FS-III: dissimilarity, energy, entropy, homogeneity; FS-IV: energy, entropy, homogeneity; FS-V: entropy, homogeneity.

FS	Linear				rbf			
	Accuracy	F1-score	Precision	Recall	Accuracy	F1-score	Precision	Recall
I	83.33	83.36	84.33	83.33	71.67	70.67	73.14	71.67
II	100	100	100	100	98.33	98.32	98.34	71.67
III	76.67	76.95	78.33	76.67	83.33	83.22	84.44	83.33
IV	36.66	21.48	30.85	76.67	53.33	40.63	35.38	53.33
V	46.67	30.08	22.40	46.6	46.67	34.71	33.03	46.67

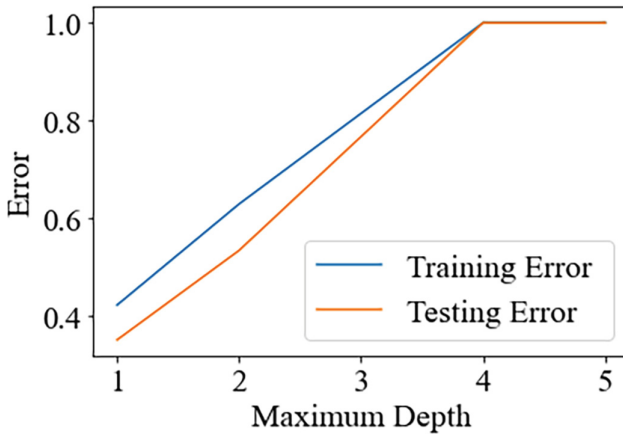


Fig. 4. Plot of training and testing error as a function of maximum depth.

5 Potential Scope for Improvement

In order to develop a suitable machine learning model, the following potentials for improvements are summarized as follows:

- **Data sufficiency:** The challenge faced in the present analysis is related to the sufficient dataset. There is no common consensus about the number of data instances sufficient for effective model performance. The performance of the model suffers unless the optimal instances are included. For the present study, extensive experiments with diverse specimens for the comprehensive dataset is indeed desirable.
- **Preprocessing Techniques:** Data preprocessing is another critical aspect of understanding the model performance. The speckle image is essentially an intensity pattern representing the surface irregularity, the scope of comparison

Table 4. Cross-validation accuracy for Random Forest classifier for the different ratios of the train-test split.

Cross-Validation Accuracy(%)		
Data Split (Train-Test)	k-fold= 5	k-fold= 10
80:20	98.75	96.25
70:30	92.85	96.42
60:40	98.33	99.16

Table 5. The Nested Cross-Validation accuracy for kNN and SVM model. All values are represented in %.

Feature Subset	Methods		
	kNN	SVM(linear)	SVM(rbf)
II	97.86	98.00	97.5
IV	96.42	-	-
V	97.86	-	-

of the results based on pre-processing is a crucial aspect to understand the model's robustness.

- **Data Augmentaion:** The process of data augmentation in its capacity of data alternation by applying random transformations to the existing data affects the efficiency of the model. It is interesting to analyze the effect on model performance by applying the data augmentation process with the speckle images.
- **Speckle Statistics:** The speckle pattern of the machined samples is sensitive to the machining operations (grinding, milling, turning) imparted to the sample during the manufacturing. The surface operation affects the speckle statistics such as speckle size and structure. With the comprehensive dataset including mixed patterns, the application of suitable machine learning configurations with various feature extraction tools can aid in the efficient prediction of sample class and surface roughness measurement.

6 Conclusion and Future Work

The classification of the surface roughness based on the speckle images of the metal specimen using state-of-the-art machine learning techniques is discussed. The performance is each technique discussed is based on the k-fold accuracy and the other performance indices such as precision, recall, and F1-score. The analysis is promising for the application of machine learning to the dataset of the speckle images encoded with the surface texture information, particularly the surface roughness. The results presented can be useful for the exploration of diverse machine learning and deep learning techniques to the high dimensional dataset for the classification and prediction of the surface roughness of

the machined samples based on the speckle images. The performance of the k-nearest neighbor is promising for the combination of correlation, dissimilarity, and homogeneity and also with correlation, dissimilarity, energy, entropy, and homogeneity. The cross-validation accuracy is 99% and nested cross-validation is 97.86% for $k = 3$ in both cases. Similarly, a Support Vector Machine with a linear kernel provides nested cross-validation is 98% for feature subsets of correlation, dissimilarity, energy, entropy, and homogeneity. The performance of Decision Tree suffers from overfitting due to small datasets. Both Decision Tree and Random Forests are expected to perform well with large datasets. The read of comparative analysis offers a strong basis for the practical *in-situ* implementation of machine learning techniques for speckle image-based surface roughness classification of the machined samples.

Based on the present analysis, future work would focus on the evaluation of the model's efficiency, particularly on a large dataset (balanced and unbalanced) containing the mixed speckle pattern of the specimens polished by various machining operations. In addition, the optimization of feature extraction techniques, investigation of the advanced algorithms, and inclusion of ensemble techniques for refining the accuracy and robustness of the model on the comprehensive dataset shall be addressed in the future.

References

1. A.L.P.Camargo, M.R.B.Dias, M.R.Lemos, M.M.Mello, L. da Silva, P.A.M.dos Santos, J.A.O.Huguenin: Estimation of statistical properties of rough surface profiles from the hurst exponent of speckle patterns. *ppl. Opt* **59**, 5957–5966 (2020)
2. Baradit, E., Gatica, C., Yáñez, M., Figueroa, J.C., Guzmán, R., Catalán, C.: Surface roughness estimation of wood boards using speckle interferometry. *Opt. Lasers Eng.* **128**, 106009 (2020). <https://doi.org/10.1016/j.optlaseng.2020.106009>
3. Bengio, Y., Delalleau, O., Le Roux, N.: The curse of dimensionality for local kernel machines. *Techn. Rep* **1258**(12), 1 (2005)
4. Chebrolu, V., Koonu, R., Raju, R., et al.: Automated evaluation of surface roughness using machine vision based intelligent systems. *Journal of Scientific & Industrial Research* **82**(1), 11–25 (2022)
5. Chen, W., Zou, B., Li, Y., Huang, C.: A study of a rapid method for detecting the machined surface roughness. *The International Journal of Advanced Manufacturing Technology* **117**, 3115–3127 (2021)
6. Connors RW, H.C.: A theoretical comparison of texture algorithms. *IEEE Trans Pattern Anal Mach Intell.* **2**(3):204–22, 110–118 (Mar 1980). 10.1109/tpami.1980.4767008
7. Corrêa, R.D.,Meireles, J.B.Huguenin J, Caetano D.P, Silva L: Fractal structure of digital speckle patterns produced by rough surfaces. *Physica A: Statistical Mechanics and its Applications* **392**, 869–874 (02 2013). 10.1016/j.physa.2012.10.023
8. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995). <https://doi.org/10.1007/BF00994018>
9. Kayahan, E., Oktem, H., Hacizade, F., Nasibov, H., Gundogdu, O.: Measurement of surface roughness of metals using binary speckle image analysis. *Tribol. Int.* **43**(1), 307–311 (2010). <https://doi.org/10.1016/j.triboint.2009.06.010>

10. Ghosh, A.K.: On optimum choice of k in nearest neighbor classification. *Computational Statistics & Data Analysis* **50**(11), 3113–3123 (2006)
11. Haralick, R.M., Shanmugam, K., Dinstein: Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**(6), 610–621 (1973). [10.1109/TSMC.1973.4309314](https://doi.org/10.1109/TSMC.1973.4309314)
12. Hearst, M., Dumais, S., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. *IEEE Intelligent Systems and their Applications* **13**(4), 18–28 (1998). <https://doi.org/10.1109/5254.708428>
13. Hitoshi Fujii, T.A.: Roughness measurements of metal surfaces using laser speckle. *J. Opt. Soc. Am.* **67**(9), 1171–1176 (1977)
14. Hurden, A.: Vibration mode analysis using electronic speckle pattern interferometry. *Optics & Laser Technology* **14**(1), 21–25 (1982)
15. Jeyapooan, T., Murugan, M.: Surface roughness classification using image processing. *Measurement* **46**(7), 2065–2072 (2013)
16. J.W.Goodman: Some fundamental properties of speckle*. *Journal of the Optical Society of America* (1917-1983) **66**(11), 1145–1150 (Nov 1976)
17. J.W.Goodman: *Speckle Phenomena in Optics Theory and Applications* Second Edition (2014)
18. Lehmann, P.: Surface-roughness measurement based on the intensity correlation function of scattered light under speckle-pattern illumination. *Appl. Opt.* **38**(7), 1144–1152 (1999). <https://doi.org/10.1364/AO.38.001144>
19. Liu, H., Zhang, S., Zhao, J., Zhao, X., Mo, Y.: A new classification algorithm using mutual nearest neighbors. In: 2010 Ninth International Conference on Grid and Cloud Computing. pp. 52–57. IEEE (2010)
20. Patil, S.H., Kulkarni, R.: Surface roughness measurement based on singular value decomposition of objective speckle pattern. *Opt. Lasers Eng.* **150**, 106847 (2022). <https://doi.org/10.1016/j.optlaseng.2021.106847>
21. Song, Y.Y., Ying, L.: Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**(2), 130 (2015)
22. Suhail, S.M., Ali, J.M., Jailani, H.S., Murugan, M.: Vision based system for surface roughness characterisation of milled surfaces using speckle line images. In: *IOP Conference Series: Materials Science and Engineering*. vol. 402, p. 012054. IOP Publishing (2018)
23. Taunk, Kashvi De, Sanjukta Verma, Srishti Swetapadma, Aleena: A brief review of nearest neighbor algorithm for learning and classification. In: 2019 international conference on intelligent computing and control systems (ICCS). pp. 1255–1260. IEEE (2019)
24. Tsai, D.M., Tseng, C.F.: Surface roughness classification for castings. *Pattern Recogn.* **32**(3), 389–405 (1999)
25. Wang, Y., Cao, J., Xu, C., Cheng, Y., Cheng, X., Hao, Q.: Moving target tracking and imaging through scattering media via speckle-difference-combined bispectrum analysis. *IEEE Photonics J.* **11**(6), 1–14 (2019)
26. Wu, X, Kumar V: *The Top Ten Algorithms in Data Mining* 1st Edition (2009)



Model Selection with a Shapelet-Based Distance Measure for Multi-source Transfer Learning in Time Series Classification

Jiseok Lee^(✉) and Brian Kenji Iwana^(iD)

Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan

jiseok.lee@human.ait.kyushu-u.ac.jp, iwana@ait.kyushu-u.ac.jp

Abstract. Transfer learning is a common practice that alleviates the need for extensive data to train neural networks. It is performed by pre-training a model using a source dataset and fine-tuning it for a target task. However, not every source dataset is appropriate for each target dataset, especially for time series. In this paper, we propose a novel method of selecting and using multiple datasets for transfer learning for time series classification. Specifically, our method combines multiple datasets as one source dataset for pre-training neural networks. Furthermore, for selecting multiple sources, our method measures the transferability of datasets based on shapelet discovery for effective source selection. While traditional transferability measures require considerable time for pre-training all the possible sources for source selection of each possible architecture, our method can be repeatedly used for every possible architecture with a single simple computation. Using the proposed method, we demonstrate that it is possible to increase the performance of temporal convolutional neural networks (CNN) on time series datasets.

Keywords: Transfer Learning · Time Series Classification · Transferability Estimation

1 Introduction

Neural networks have widespread usage in time series recognition. For example, temporal Convolutional Neural Networks (CNN) [19] have been shown to be effective across many time series domains [2, 38]. However, often, neural networks require large amounts of data [10, 15]. Also, acquiring large amounts of annotated data can take time and effort.

Several ideas exist to solve the problem of the requirement of large amounts of annotated data, such as transfer learning, self-supervised learning, data augmentation, etc. In particular, transfer learning has become a popular method of

This work was partially supported by MEXT-Japan (Grant No. 23K16949).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPD 2024, LNCS 15327, pp. 160–175, 2025.
https://doi.org/10.1007/978-3-031-78398-2_11

initializing neural networks. In transfer learning, to alleviate the need for data, neural networks can be trained on larger *source* datasets and fine-tuned for *target* datasets. In this way, the weights of the neural network can be trained to extract generalized features [46] and be used for the target task. In the image recognition domain, transfer learning is a standard practice. For example, using pre-trained models trained with ImageNet [8] in image recognition is standard practice. However, for time series, transfer learning is still a budding field [10].

In order to realize an effective method of transfer learning for time series, we propose a combination of multi-source transfer learning with a novel shapelet-based distance measure used for dataset selection. Specifically, to increase the effectiveness of transfer learning and the source dataset’s size, we propose a method of using multiple datasets for pre-training.

However, selecting the datasets for transfer learning is complex. Fawaz et al. [10] demonstrated that the choice of dataset for transfer learning for time series has a large effect on the effectiveness of transfer learning. Notably, only some datasets increased the accuracy of the model. Oftentimes, the accuracy was decreased when using an inappropriate dataset.

Thus, we propose a dataset distance-based measure to select the appropriate datasets for our multi-source pre-training. To do this, first, we extract discriminative shapelets using shapelet discovery [42]. Next, a dataset distance measure is created by comparing the discriminative shapelets between the source and target datasets. The idea is that datasets with similar discriminative shapelets would have similar features, thus leading to more effective transfer learning.

The contribution of this paper is as follows:

- We propose a new method of predicting and selecting source datasets for transfer learning in temporal neural networks. This method uses extracted shapelet similarity between the target and possible source datasets.
- We create a transfer learning method that combines multiple sources into one super dataset.
- We evaluate the proposed method on all 128 time series datasets from the 2018 UCR Time Series Archive (UCR Archive) [7].
- We provide code for easy transfer learning at <https://github.com/uchidalab/time-series-transferability>

2 Related Works

2.1 Transfer Learning for Time Series

Transfer learning has been used for various applications in image recognition [46]. Furthermore, transfer learning has become the standard practice for training networks, as pre-trained weights are available for the most popular image recognition network architectures.

Conversely, transfer learning is less common for time series recognition and temporal neural networks [10] outside of Natural Language Processing (NLP).

However, there have been a few works that demonstrate the usefulness of transfer learning in the time series domain [10, 36, 39]. Other examples include using transfer learning with health data [4], human activity recognition [1], prediction of internet load [9], and fall detection [23]. De Souza et al. [33] proposed decomposing time series into shapelets and noise and using the decompositions to pre-train models. In comparison to our method, we use shapelets as a distance measure between datasets and not directly to train models.

For the source selection, several works demonstrated that source selection with dataset similarity, computed by using Dynamic Time Warping (DTW) [27] distance, can be effective [10, 22, 43].

2.2 Multi-Source Transfer Learning

There have been a few works that use multiple source datasets for transfer learning. For example, Yao and Doretto [41] extend TrAdaBoost [6], a method of boosting transfer learning, to use with multiple sources. Huang et al. [13] improve on this and propose SharedBoost for multiple source transfer learning. Multi-transfer [34] combines multi-view and multi-source transfer learning. For multi-source transfer learning, Song et al. [32] use the conditional probability difference to weight source domains.

Multi-source transfer learning has also been used for time series data. One of the typical methods of multi-source transfer learning is to use a preliminary classifier to select the sources. For example, for electroencephalogram (EEG) data, Jinpeng Li et al. [20] trained each source individually, and then they tested the target domain on each and selected the top-performing models. Ren et al. [26] classify EEG data using a multi-source model. Huiming Lu et al. [22] use an ensemble model to implement multi-source transfer learning for building energy prediction.

Also, Yao et al. [40] use multi-source transfer learning with Variational Mode Decomposition to improve PM2.5 concentration forecasting while selecting sources using Euclidean Distance and Maximum Mean Discrepancy. Lotte and Guan [21] use a search algorithm to combine different datasets. Senanayaka et al. [29] used a similarity-based approach for multi-source transfer learning to generate a mixed domain of multiple sources and targets in the pre-training stage.

3 Transferability Measure

3.1 Problem Setting

Given source dataset $\mathcal{S} = \{(\mathbf{s}_1, z_1), \dots, (\mathbf{s}_m, z_m), \dots, (\mathbf{s}_M, z_M)\}$, where (\mathbf{s}_m, z_m) is the m -th pair of pattern \mathbf{s}_m and respective label z_m , transfer learning aims to train a network f with \mathcal{S} , so that it will be a practical starting point for target dataset $\mathcal{T} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n), \dots, (\mathbf{x}_N, y_N)\}$, where (\mathbf{x}_n, y_n) is the n -th pair of pattern \mathbf{x}_n and respective label y_n . Unlike domain adaptation, there is no assumption that the task of the source and target datasets are related.

As Fawaz et al. [10] found, not all source datasets are useful for transfer learning with time series. Thus, a suitable source dataset \mathcal{S} for each target dataset \mathcal{T} should be determined. Under the problem setting, this determination should be performed before training \mathcal{T} , i.e., without exhaustively fine-tuning \mathcal{T} on all possible datasets.

Two types of measures have been proposed to solve the source selection problem. One class of measures is to estimate the transferability of the pre-trained network f , and the other is to measure the similarity of the datasets.

3.2 Transferability Estimation

Transferability estimation measures attempt to predict how effective transfer learning will be for model f , pre-trained on source dataset \mathcal{S} , for target dataset \mathcal{T} . These methods first use models f pre-trained on datasets \mathcal{S} to predict target dataset \mathcal{T} . Specifically, prediction $f(x_n)$ is done using the data x_n of \mathcal{T} with the source labels $c \in C$ and the features $\mathcal{F}(x_n) \in \mathcal{F}$ are extracted from model f trained by \mathcal{S} . While the source labels C might be unrelated to the target task, the outputs are used for the transferability estimation. In other words, models trained for \mathcal{S} are used as-is with dataset \mathcal{T} , and the amount of information inferred by the model is measured and used as a transferability estimation measure. Some transferability estimation measures include, Log Expected Empirical Prediction (LEEP) [24], Negative Conditional Entropy (NCE) [37], Log Maximum Evidence (LogME) [45], Transrate [12], and H-score [3].

For example, LEEP [24] first predicts the target dataset \mathcal{T} using trained f . LEEP is then calculated by:

$$\text{LEEP}(\mathcal{S}, \mathcal{T}) = \frac{1}{N} \sum_{n=1}^N \log \left(\sum_{c \in C} \hat{P}(y_n|c) f(x_n) \right), \quad (1)$$

where $\hat{P}(y_n|c)$ is the empirical conditional distribution calculated by:

$$\hat{P}(y_n|c) = \frac{\hat{P}(y_n, c)}{\hat{P}(c)} = \frac{\frac{1}{N} \sum_{n: y_n=c} f(x_n)}{\frac{1}{N} \sum_{n=1}^N f(x_n)}. \quad (2)$$

LEEP is the average log-likelihood of the prediction of \mathcal{T} in trained network f multiplied by $\hat{P}(y_n|c)$ for each source class c .

Dataset Similarity Measure for Source Selection As an alternative to transferability estimation, dataset similarity can also be used to predict transferability. The previous methods are helpful because only the pre-trained network f and not the original dataset \mathcal{S} is needed to calculate transferability. However, unlike image recognition, standard models with downloadable weights for time series recognition are lacking. Therefore, requiring pre-trained networks is a detriment because it requires training many networks before pre-training the actual network for the task.

Conversely, measuring the distance between datasets only requires access to the datasets. Following this, Fawaz et al. [10] proposed to use DTW [27] between representative time series patterns from each class in the target and source datasets. The representative time series is the average time series of each class found by DTW Barycenter Averaging (DBA) [25]. They defined the distance between datasets as the distance between the most similar average time series from each dataset. In this paper, we define this method as DBA-DTW. By using the dataset-based distance measure, the appropriate source dataset can be selected for the target dataset. The benefit of this and the proposed method is that no initial trained model is required to measure transferability.

4 Multi-Source Transfer Learning

We propose a simple yet effective method of combining multiple source datasets for transfer learning. As shown in Fig. 1, to perform multi-source pre-training, we compile multiple source datasets $\mathcal{S}_1, \dots, \mathcal{S}_i, \dots, \mathcal{S}_I$ into one super dataset $\mathcal{S}_{\text{Multi}} = \{\mathcal{S}_1, \dots, \mathcal{S}_i, \dots, \mathcal{S}_I\}$. In order to transfer knowledge from a model trained with multiple sources, the datasets are pre-processed so that they have the same number of time steps as the target dataset. Note that by resampling the source datasets, the features and characteristics of the datasets might not be preserved. However, this fact is optional because the purpose of the pre-training is to acquire a robust set of initial weights for transfer learning and not the classification accuracy of the source datasets. In addition to resampling, $\mathcal{S}_{\text{Multi}}$ is balanced so that each sub-dataset has the same number of time series. To balance $\mathcal{S}_{\text{Multi}}$, oversampling is performed while preserving the class ratios. This is done due to the discrepancy in the size of possible source datasets; it ensures that every dataset has an equal contribution to the transfer learning.

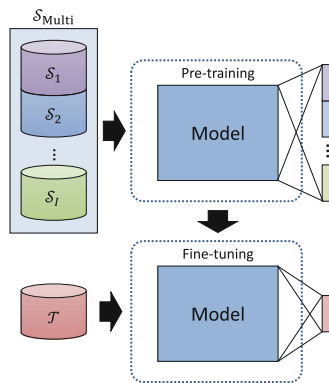


Fig. 1. An illustration of our multi-source transfer learning. Source datasets \mathcal{S}_i are selected using a transferability measure, and the neural network is pre-trained. The trained weights are then fine-tuned using target dataset \mathcal{T} .

As shown in Fig. 1, in order to train a network with multiple datasets, the one-hot vector of the ground truth of each source dataset \mathcal{S}_i are concatenated, and the output nodes are extended accordingly. In this way, the network is then trained using $\mathcal{S}_{\text{Multi}}$ for a classification task using all of the classes from all of the source datasets. The result is the ability to pre-train a network with a larger dataset with a larger number of classes.

After training the network, fine-tuning can be performed as typical transfer learning does. The weights of the trained network can be used as an initialization for a target dataset and fine-tuned for a specific task. While the experimental results use temporal CNNs, there is no theoretical limitation on the type of neural network used.

5 Shapelet Similarity-based Source Selection

In order to use the proposed transfer learning method effectively, some source datasets need to be selected. However, as mentioned, selecting the source datasets needs to be performed. Thus, we propose a new method of measuring the transferability of networks through a novel dataset similarity measure using discriminative shapelets.

5.1 Shapelet

A shapelet refers to a subsequence extracted from time series data that are maximally representative of a class [42]. These subsequences are intended to encapsulate fundamental patterns or discriminative features within a class. For example, the circled subsequences in Fig. 2 are well discovered within a class and represent differences between the two classes. In the figure, the circled shapelets are segments of the time series unique to each class.

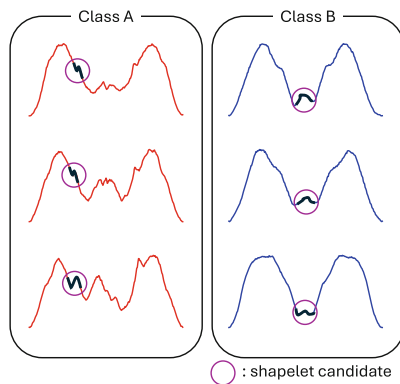


Fig. 2. Examples of shapelets from the Arrowhead dataset. The left and right figures are three time series patterns from the same classes.

5.2 Matrix Profile for Shapelet Discovery

Since a shapelet can be any subsequence from a time series, finding the maximum representative shapelet would be too costly with brute force. In order to overcome this issue, Matrix Profile [44] has been shown to find discriminative shapelet candidates efficiently.

Matrix Profile is an algorithm that represents a time series based on the distances between subsequences of that time series and their nearest neighbor. Specifically, given an ordered list of all subsequences \mathcal{A} of a single time series \mathbf{t} , Matrix Profile \mathbf{p} is a sequence that holds the distances between each subsequence \mathcal{A}_r to its nearest neighbor, or:

$$\mathbf{p} = \|\mathcal{A}_1 - \mathcal{E}_1\|, \dots, \|\mathcal{A}_r - \mathcal{E}_r\|, \dots, \|\mathcal{A}_R - \mathcal{E}_R\|, \quad (3)$$

where \mathcal{E}_r is the nearest subsequence of \mathcal{A} to the respective \mathcal{A}_r and $\|\cdot\|$ is the sum of the pair-wise Euclidean distances between each element in the subsequences. By finding Matrix Profile \mathbf{p} using time series \mathbf{t} , Matrix Profile can be used as a fast motif and discord discovery method.

In order to use Matrix Profile for shapelet discovery, a few modifications are performed. First, given a dataset \mathcal{S} , all of the time series of each class c are concatenated into a single time series $\mathbf{t}^{(c)}$. For example, given class 1 and class 2, a time series $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$ are created. Then, instead of just calculating Matrix Profile using the nearest neighbors of $\mathcal{A}^{(1)}$ with itself in (3), a Matrix Profile calculation is made for each combination of the two classes, or $\mathbf{p}^{(1,1)}$, $\mathbf{p}^{(1,2)}$, $\mathbf{p}^{(2,2)}$, and $\mathbf{p}^{(2,1)}$. Finding the highest values in the differences $\mathbf{p}^{(1,2)} - \mathbf{p}^{(1,1)}$ and $\mathbf{p}^{(2,1)} - \mathbf{p}^{(2,2)}$ will identify the maximally representative shapelet candidates for class 1 and 2, respectively. Because this is only compatible with two-class classification, we extend shapelet discovery via Matrix Profile using a one-versus-all approach for each class.

5.3 Shapelet Similarity-based Source Selection

Now that the representative shapelets $\mathcal{P}^{(S)}$ and $\mathcal{P}^{(T)}$ can be found for each source dataset \mathcal{S} and target dataset \mathcal{T} , respectively, we use them as a basis for a dataset distance measure. We propose two shapelet distance measure schemes, Average Shapelet, and Minimum Shapelet distances. Fig. 3 represents an overview of Average Shapelet and Minimum Shapelet.

Average Shapelet takes the average distance for each combination of $\mathcal{P}_i^{(S)}$ and $\mathcal{P}_j^{(T)}$, or:

$$D_{\text{as}} = \frac{1}{IJ} \sum_{i,j} \|\mathcal{P}_i^{(S)} - \mathcal{P}_j^{(T)}\|, \quad (4)$$

where $\mathcal{P}_i^{(S)}$ and $\mathcal{P}_j^{(T)}$ are the i -th and j -th shapelet in $\mathcal{P}^{(S)}$ and $\mathcal{P}^{(T)}$, respectively. By using the average distance between shapelets, this measure compares all of the discriminative features of the datasets simultaneously. The general idea of

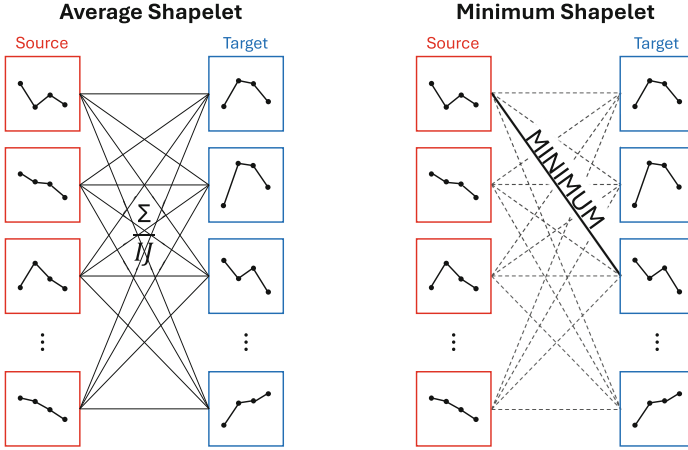


Fig. 3. Overview of Average Shapelet and Minimum Shapelet.

this measure is that if all of the shapelets of the datasets are similar, then the datasets might be similar.

Minimum Shapelet is defined as the distance between the most similar pair of shapelets of \mathcal{S} and \mathcal{T} , or:

$$D_{\text{ms}} = \min_{i,j} \|\mathcal{P}_i^{(\mathcal{S})} - \mathcal{P}_j^{(\mathcal{T})}\|. \quad (5)$$

Instead of measuring all of the features, this measure allows the distance calculation to ignore features that might be specific to a dataset.

6 Experimental Result

6.1 Dataset

The experiments were conducted using all of the UCR Archive [7], which consists of 128 datasets. We use the predetermined training and test set split provided by the archive. Also, no pre-processing was performed except for resizing datasets through Gaussian smoothing with different lengths.

6.2 Settings and Architecture

For the experiment, we adopted a 1-dimensional CNN model based on the VGG architecture [31]. The convolutional network used three blocks of convolutional layers and a pooling layer. The first block has two convolutional layers of size 3, and the subsequent blocks have three convolutional layers. Max pooling is used with the first two blocks, and global average pooling (GAP) is used with the final block. While GAP is not required for the proposed method, it is required

for LEAP, NCE, H-score, Transrate, and LogMe due to having different-sized datasets; thus, we used it for all evaluations.

For training, we pre-train the network for 10,000 iterations with Adaptive Moment Estimation (Adam) optimizer [17] with an initial learning rate of 0.0001. For transfer learning, we then fine-tune the network for 5,000 iterations. The batch size is set to 32 for both pre-training and fine-tuning. For statistical validity, we fine-tuned the model three times in order to have the mean of the three models' performances. Since the traditional transferability measures, LEEP, NCE, LogME, Transrate, and H-Score, require a trained network, an initial network to calculate the transferability is trained for 5,000 iterations.

There are two hyperparameters associated with the shapelet discovery by Matrix Profile. First, we use a fixed shapelet size of 15 because this is the largest possible size on the UCR Archive's smallest dataset. Next, we use the top 10 shapelet candidates per class.

6.3 Comparative Evaluation

To evaluate the proposed method, we compared it to not using transfer learning, to using transfer learning using a dataset selected by a shapelet-based distance measure, and to using a dataset selected by the other transferability metrics. The comparative measures used for source selection include using DTW between DBA class representatives (DBA-DTW) [10], LEEP [24], NCE [37], Transrate [12], LogME [45], and H-score [3].

Table 1. Average test accuracy.

Method	Accuracy (%)
No Transfer Learning (TL)	74.18
TL w/ Random Source	76.89
TL w/ DBA-DTW	78.85
TL w/ H-Score	77.24
TL w/ LEEP	76.43
TL w/ LogME	79.59
TL w/ NCE	78.92
TL w/ Transrate	77.22
Proposed w/ Average Shapelet (10 shapelets)	79.91
Proposed w/ Minimum Shapelet (10 shapelets)	80.25

The results of the experiments are shown in Table 1. In the table, TL represents typical transfer learning that selects a single source dataset, and MTL represents our proposed method, Multi-source Transfer Learning. Compared to the model with random initialized weights, Multi-source Transfer Learning with a Minimum Shapelet of 10 candidates showed the highest improvement. On the

other hand, the Average Shapelet showed lower improvement than the Minimum Shapelet; however, it is still better than other comparative methods.

Additionally, we looked into the datasets that showed better and worse performance by adopting our proposed method. Fig. 4 shows sample plots of the datasets. From the sample plot, we can affirm that our proposed method works better with datasets with more clear features. Also, the datasets with more noise showed worse performance than our proposed method. This trend is due to our proposed method focusing on shapelet similarity.

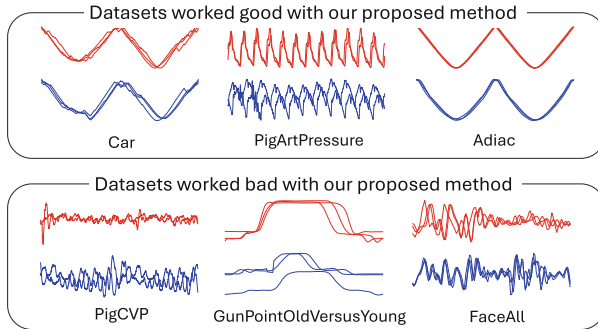


Fig. 4. Sample plot of datasets of UCR Archive. Random three samples are plotted from two classes. The upper three datasets performed better with our proposed method, and the lower three datasets showed adverse effects by adopting our proposed method.

To compare the proposed method to other methods found in the literature, we compare it against reported results on the UCR Archive from methods that use neural networks. Fig. 5 is a Nemenyi post-hoc test diagram comparing the proposed methods, the base models, and various reported evaluations. The comparison models include a temporal Residual Network (ResNet) [38], Fully Convolutional Network (FCN) [38], MLP [38], Multi-scale CNN (MCNN) [5], Time Warping Invariant Echo State Network (TWI-ESN) [35], Time Le-Net (t-LeNet) [18], universal Encoder [30], LSTM [11], BLSTM [28], LSTM-FCN [16]. The models were evaluated by Wang et al. [38], Ismail Fawaz et al. [14], and Iwana and Uchida [15]. The figure shows that the networks used by the proposed method are comparable to the other state-of-the-art neural networks on the same datasets.

7 Discussion

7.1 Ablation Study

We compared the classification performances with multi-source pre-training with shapelet-based source selection, multi-source pre-training with random source

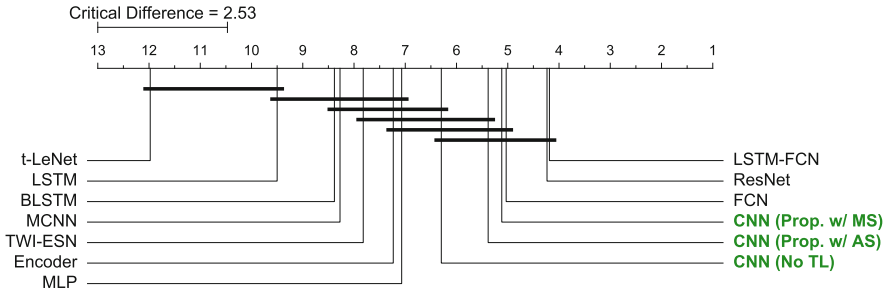


Fig. 5. A Nemenyi post-hoc test diagram. The proposed methods are in green. The numbers indicate the average rank when tested on all 128 datasets.

selection, and without transfer learning. Fig. 6 compares the performances of each dataset of the UCR Archive. Our proposed method showed significantly better results than the random initialized model and multi-source transfer learning with random source selection. Specifically, the result on the right figure demonstrates that our proposed shapelet similarity-based source selection is effective for source selection on multi-source transfer learning. Also, according to the t-test, our proposed method is effective with most datasets, $p < 0.001$.

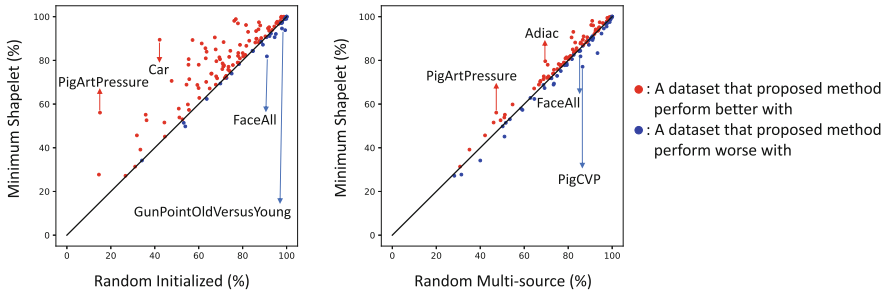


Fig. 6. Ablation study of our proposed method. The Y-axis represents the classification accuracy of our proposed method with 14 sources selected by Minimum Shapalet. The X-axis of the left and right figures represent the classification accuracy of random initialized models and proposed multi-source transfer learning with 14 random sources.

7.2 Number of Datasets

In order to examine how the hyperparameters for the proposed method affect performance, we examined the number of shapelet candidates, the number of source datasets, and the two shapelet similarity-based distance measures. For the number of shapelet candidates, we discovered three, five, and ten shapelet

candidates for every dataset. For the number of source datasets, we selected 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 datasets according to shapelet-based distance measures. Finally, we examined the Average Shapelet and the Minimum Shapelet for the shapelet-based distance measure.

The experimental results in Fig. 7 showed better performance with more sources for all selection methods, including Random multi-source selection. Thus, increasing the number of datasets using our multi-source transfer learning method effectively increases the effect of pre-training while alleviating the risk of negative transfer. However, there was no significant difference when the number of datasets was more than 10. This implies a diminishing returns effect with the number of datasets. Thus, there is a limit to how many datasets should be combined.

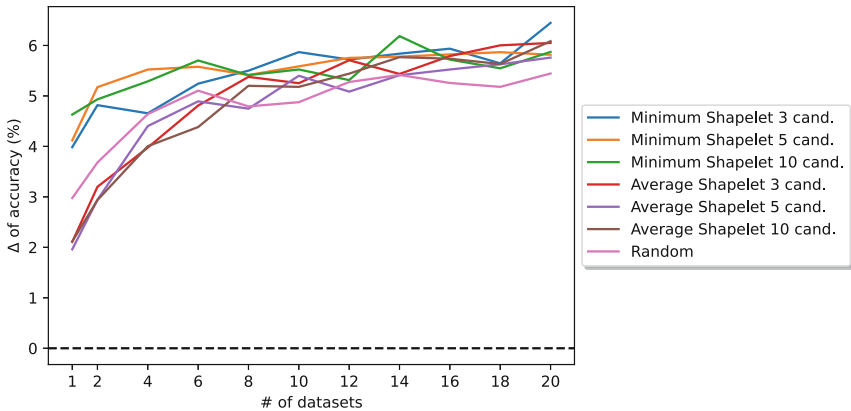


Fig. 7. Experimental result of examining hyperparameters.

However, one possible reason for the diminishing returns is due to our training scheme. We set the number of pre-training iterations to a fixed number, no matter how many datasets and data samples were added for a fair comparison. Therefore, it might be possible to increase the accuracy further using multiple datasets with more iterations.

7.3 Similarity Based Source Selection

Generally, with a small-scale target dataset, pre-training with a closely related source dataset allows for more efficient training while reducing the risk of overfitting. However, many time series tasks such as EGG, Speeches, and Gestures have different features and usually contain a small number of patterns. Thus, selecting a source dataset for time series transfer learning is often a big challenge, and transfer learning with non-related tasks is usually not helpful.

According to the experimental result, as depicted in Table 1, selecting a source dataset based on dataset ranking metrics of DBA-DTW and Minimum

Shapelet resulted in superior performance compared to a random source selection and even to the classic transferability measures. Therefore, for time series classification, where lack of data is a frequent challenge, measuring time series similarity can serve as a valuable indicator of transferability.

7.4 Computational Time

Regarding computational time, shapelet similarity-based source selection has a huge benefit compared to other transferability estimation metrics. Shapelet similarity-based dataset ranking has two steps of calculation: one step is to generate shapelet, and the other is to calculate the similarity among generated shapelets. Thus, the shapelet similarity-based source selection requires $O(n^2 + w)$; $O(n^2)$ to generate shapelet and $O(w_s)$ to calculate the similarity among shapelets, where n is the length of the dataset and w_s is the length of shapelet. The DBA-DTW also has the benefits of computational time compared to the other transferability estimation measures; however, it is not like the shapelet similarity-based method. DBA-DTW requires $O(i \cdot n^2 + w_d^2)$; $O(i \cdot n^2)$ to generate prototypes through DBA and $O(w_d^2)$ to calculate distance among generated shapelets, where w_d is the length of a prototype of DBA.

The other transferability estimation measures, such as LEEP, require a pre-trained model to measure transferability. Training a model requires significant time, and they have a disadvantage in that they need to re-calculate the transferability when the model architecture is changed. In this research, it is required to train 128×128 models for evaluation. However, our proposed method requires no re-calculation even though the model architecture has been changed. Thus, unlike transferability estimation measures, additional datasets can be calculated quickly and used for other tasks.

8 Conclusion

In this paper, we suggest using transfer learning for temporal neural networks using a proposed multi-source pre-training. Specifically, we demonstrate that by combining multiple datasets into a super dataset using pre-processing and adjusting the classification task with the concatenation of the classes, it is possible to pre-train a network using a large amount of data and classes.

Furthermore, in order to select appropriate datasets out of the large number of possible datasets that exist, we propose a new transferability measure based on shapelets. Our novel method calculates the distance between datasets using a shapelet similarity. The shapelet-based distance compares the class-discriminative shapelets between classes of the target dataset and the source dataset. We demonstrate that by using multi-source transfer learning with our shapelet similarity-based source selection, it is possible to increase the time series classification accuracy with little downside.

In future work, we will investigate the combinations of source datasets to optimize our proposed method further. We hope to contribute to the time series classification community by continuing to expand upon these techniques.

References

1. An, S., Bhat, G., Gumussoy, S., Ogras, U.: Transfer learning for human activity recognition using representational analysis of neural networks. *ACM Transactions on Computing for Healthcare* **4**(1), 1–21 (2023)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint [arXiv:1803.01271](https://arxiv.org/abs/1803.01271) (2018)
3. Bao, Y., Li, Y., Huang, S.L., Zhang, L., Zheng, L., Zamir, A., Guibas, L.: An information-theoretic approach to transferability in task transfer learning. In: *ICIP* (2019)
4. Clark, R., Doyle, T.E.: A priori quantification of transfer learning performance on time series classification for cyber-physical health systems. In: *IEEE CCECE* (2022)
5. Cui, Z., Chen, W., Chen, Y.: Multi-scale convolutional neural networks for time series classification. arXiv preprint [arXiv:1603.06995](https://arxiv.org/abs/1603.06995) (2016)
6. Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Boosting for transfer learning. In: *ICML* (2007)
7. Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., Hexagon-ML: The ucr time series classification archive (October 2018), https://www.cs.ucr.edu/~eamonn/time_series_data_2018/
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *CVPR*. pp. 248–255 (2009)
9. Dridi, A., Afifi, H., Moun gla, H., Boucetta, C.: Transfer learning for classification and prediction of time series for next generation networks. In: *IEEE ICC* (2021)
10. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Transfer learning for time series classification. In: *IEEE ICBD* (2018)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
12. Huang, L.K., Huang, J., Rong, Y., Yang, Q., Wei, Y.: Frustratingly easy transferability estimation. In: *ICML*. pp. 9201–9225 (2022)
13. Huang, P., Wang, G., Qin, S.: Boosting for transfer learning from multiple data sources. *Pattern Recogn. Lett.* **33**(5), 568–579 (2012)
14. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. *Data Min. Knowl. Disc.* **33**(4), 917–963 (2019)
15. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE* (2021)
16. Karim, F., Majumdar, S., Darabi, H., Chen, S.: Lstm fully convolutional networks for time series classification. *IEEE Access* **6**, 1662–1669 (2017)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
18. Le Guennec, A., Malinowski, S., Tavenard, R.: Data augmentation for time series classification using convolutional neural networks. In: *ECML/PKDD WAALTD* (2016)
19. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
20. Li, J., Qiu, S., Shen, Y.Y., Liu, C.L., He, H.: Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Transactions on Cybernetics* pp. 1–13 (2019)

21. Lotte, F., Guan, C.: Learning from other subjects helps reducing brain-computer interface calibration time. In: IEEE ICASSP (2010)
22. Lu, H., Wu, J., Ruan, Y., Qian, F., Meng, H., Gao, Y., Xu, T.: A multi-source transfer learning model based on lstm and domain adaptation for building energy prediction. *Int. J. of Electrical Power & Energy Systems* **149**, 109024 (2023)
23. Maray, N., Ngu, A.H., Ni, J., Debnath, M., Wang, L.: Transfer learning on small datasets for improved fall detection. *Sensors* **23**(3), 1105 (2023)
24. Nguyen, C., Hassner, T., Seeger, M., Archambeau, C.: LEEP: A new measure to evaluate transferability of learned representations. In: ICML. pp. 7294–7305 (2020)
25. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recogn.* **44**(3), 678–693 (2011)
26. Ren, R., Yang, Y., Ren, H.: EEG emotion recognition using multisource instance transfer learning framework. In: ICICML (2022)
27. SAKOE, H., CHIBA, S.: Dynamic programming algorithm optimization for spoken word recognition. *Readings in Speech Recognition* pp. 159–165 (1990)
28. Schuster, M., Paliwal, K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997)
29. Senanayaka, A., Al Mamun, A., Bond, G., Tian, W., Wang, H., Fuller, S., Falls, T., Rahimi, S., Bian, L.: Similarity-based multi-source transfer learning approach for time series classification. *International Journal of Prognostics and Health Management* **13**(2) (2022)
30. Serra, J., Pascual, S., Karatzoglou, A.: Towards a universal neural network encoder for time series. In: AIRD. pp. 120–129 (2018)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
32. Song, H.J., Park, S.B.: Identifying intention posts in discussion forums using multi-instance learning and multiple sources transfer learning. *Soft. Comput.* **22**(24), 8107–8118 (2017)
33. de Souza, A.F.M., Cassenote, M.R.S., Silva, F.: Transfer learning of shapelets for time series classification using convolutional neural network. In: Britto, A., Valdivia Delgado, K. (eds.) *Intelligent Systems*. pp. 325–339 (2021)
34. Tan, B., Zhong, E., Xiang, E.W., Yang, Q.: Multi-transfer: Transfer learning with multiple views and multiple sources. In: SIAM ICDM (2013)
35. Tanisaro, P., Heidemann, G.: Time series classification using time warping invariant echo state networks. In: IEEE ICMLA. pp. 831–836 (2016)
36. Thompson, A.: Transfer learning with time series prediction: Review. *SSRN Electronic Journal* (2022)
37. Tran, A., Nguyen, C., Hassner, T.: Transferability and hardness of supervised classification tasks. In: ICCV (2019)
38. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: IJCNN (2017)
39. Weber, M., Auch, M., Doblender, C., Mandl, P., Jacobsen, H.A.: Transfer learning with time series data: A systematic mapping study. *IEEE Access* **9**, 165409–165432 (2021)
40. Yao, B., Ling, G., Liu, F., Ge, M.F.: Multi-source variational mode transfer learning for enhanced pm2. 5 concentration forecasting at data-limited monitoring stations. *Expert Systems with Applications* **238**, 121714 (2024)
41. Yao, Y., Doretto, G.: Boosting for transfer learning with multiple sources. In: IEEE CVPR (2010)

42. Ye, L., Keogh, E.: Time series shapelets. In: ACM KDD (2009)
43. Ye, R., Dai, Q.: Implementing transfer learning across different datasets for time series forecasting. *Pattern Recogn.* **109**, 107617 (2021)
44. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A., Keogh, E.: Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In: IEEE ICDM (2016)
45. You, K., Liu, Y., Wang, J., Long, M.: Logme: Practical assessment of pre-trained models for transfer learning. In: ICML. pp. 12133–12143 (2021)
46. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. *Proceed. IEEE* **109**(1), 43–76 (2021)



DACOA: Diffusion-Aligned Coherent Augmentation and Consistency Constraint Strategies for Federated Domain Generalization

Guangshuo Wang, Yuesheng Zhu^(✉), and Guibo Luo^(✉)

Shenzhen Graduate School, Peking University, Beijing, China
wanggsh@stu.pku.edu.cn, {zhuys, luogb}@pku.edu.cn

Abstract. Federated learning is a privacy-preserving, decentralized machine learning approach, which faces the challenge of non-identically distributed (non-iid) data between train-test data and across client domains. Previous methods generally exchange domain information between clients to perform data augmentation. While bringing privacy risks and communication costs, these methods also destroy the coherence of images. To tackle these issues, we propose Diffusion-Aligned Coherent Augmentation (DACOA), a diffusion-based and text-guided style transfer method. By composing different domains and labels as prompts, clients are guided to perform cross-domain image augmentation with high quality, thereby learning robust representation against domain shift. To better utilize the augmentation results and help the model focus on semantic information, we conduct alignment on both the feature dimensions and prediction results. We introduce the Domain Aligning Contrastive Learning (DaCon) loss, which brings the feature similarity of the same label closer. Also, we introduce the Semantic-Consistency KL (SCKL) loss, aligning the prediction results of the augmented images with the original classification results. Our model outperforms state-of-the-art FedDG methods through comprehensive experiments. What's more, we achieve 3.21%, 1.76% and 5.01% improvement on PACS, Office-Home, and Digits-DG benchmarks. Ablation study validates the efficacy of each module.

Keywords: Federated domain generalization · Latent diffusion model · Data augmentation

1 Introduction

There is significant progress in the field of deep learning, however, most of it relies on the assumptions of independent and identically distributed (IID) data. This reliance often leads to a noticeable decline in performance when these methods are tested on Out-of-Distribution (OOD) datasets. The Domain Generalization (DG) technique was introduced to address this issue, aiming to enhance the generalization to unseen data. In the DG framework, training occurs on known and

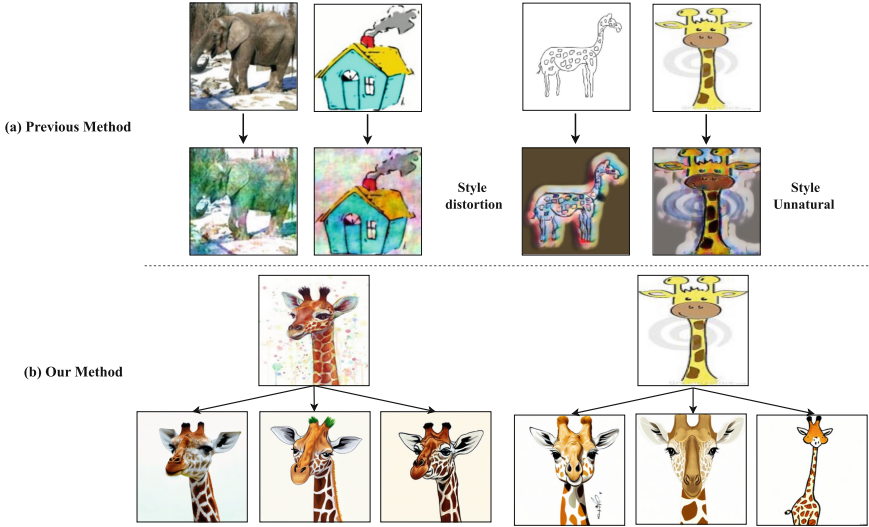


Fig. 1. Problems with previous augmentation methods.

utilized source domains, with the aim to improve performance on unknown target domains. Nevertheless, the growing popularity of distributed training has led to less sharing of inter-domain data, which makes the conventional DG method invalid.

To tackle this issue, Federated Domain Generalization (FedDG) was proposed [10], offering a decentralized and privacy-preserving training approach. In FedDG, each client’s training data serves as a source domain, characterized by unique, non-shareable, and non-IID data distributions. After training locally, each client’s model parameters are sent to a server where they are aggregated to form a new global model. This model is then used for inference in the unknown domains. However, the FedDG scenario presents its own set of challenges. Most traditional DG methods require the simultaneous integration of information from multiple domains for domain-invariant learning, making them less suitable for the more realistic and demanding FedDG scenario.

Previous methods attempted to exchange frequency domain [10] or domain statistical information [3] between clients and use it for augmentation. However, these methods often result in unnatural results and destroy the style coherence of the image. As shown in Fig. 1., frequency domain enhancement methods are prone to style distortion, while statistical information enhancement methods generate unnatural new images. Moreover, exchanging information between clients still brings privacy risks and greater communication costs. Introducing the Latent diffusion model (LDM) [12] can solve the above problems well. LDM effectively understands and maintains the characteristics of the image during the data enhancement process, so it can provide effective enhancement while maintaining the naturalness of the resulting image, thereby guiding the client model

to learn domain invariant features. In addition, LDM can use text prompts to control the generation of results, thereby achieving cross-domain enhancement without communication between domains, complying with privacy regulations, and reducing inter-domain communication.

Based on the aforementioned motivations, we propose a Diffusion-Aligned Coherent Augmentation (DACOA) method. We leverage a pretrained LDM for data augmentation, performing style transfer using text prompts like "a {domain} of {class}". We also utilize CLIP's text encoder to generate text embeddings as conditional embeddings for the LDM's denoising process. For a specific instance, we set a prompt to generate augmented data with the same label but different domains. By directly utilizing the LDM model, clients can perform style transfer without accessing data from other domains, thereby achieving a higher coverage of feature range and learning domain invariant features.

To better utilize DACOA and enhance the model's focus on semantic information for improved generalization, we aim to align the augmented data with the original data in terms of features and model prediction results. To achieve this, we propose the Domain Aligning Contrastive Learning (DaCon) loss and Semantic-Consistency KL (SCKL) loss. DaCon operates on the feature dimension, aiming to increase the similarity of features with the same label and push away features with different labels. SCKL, on the other hand, calculates the sharpened KL divergence on the predicted logits, with the goal of making the predictions of augmented data closer to those of the original data. Our main contributions can be summarized as follows:

- We propose the diffusion-aligned coherent augmentation (DACOA) method, allowing clients to generate style-transferred image without ruining privacy protocol, enabling model to learn domain-invariant feature.
- To better utilize DACOA, we incorporate domain aligning contrastive learning (DaCon) loss and semantic-consistency KL (SCKL) loss to draw augmented feature and prediction closer to the original one.
- Comprehensive experiments on several datasets demonstrate that our results achieve state-of-the-art performance, validating the effectiveness of each module.

2 Related Work

An important prerequisite for generalizing well on unseen test domains is that the model can extract domain invariant features [1] during the training process. That is to say, after model training, we hope that the features extracted by its encoder only depend on the input category information, and do not change with the input texture and style information. In traditional DG methods, the model mostly needs to use information from multiple domains to shorten the discrepancy between domains. The most commonly used method is data augmentation. [20], which belongs to image augmentation, generates new domain images from the source data by optimizing a divergence measure based on optimal transport. The feature augmentation method enhances the intermediate features of

the model[7,21]. For instance,[7] computes the statistical features of multiple domains and augmenting them through the encoder.

However, under the schema of FedDG, most conventional DG methods are no longer applicable since they need to simultaneously utilize information from multiple domains. One solution in FedDG is to optimize the training method of local models. [10] first proposed a solution for FedDG scenarios, exchanging data distribution information between clients to carry out enhancement in the continuous frequency space. Similarly, [3] achieves a consistent distribution by building a style bank, allowing clients to construct style transfer and generate new domains. FADH [16] trains the local models with domain hallucinations for robustness. Another idea is to optimize the parameter aggregation process to extract domain invariant features. COPA [15] employs batch-wise mixed normalization and aggregation, [18] optimizes the parameter aggregation process in FL, conducts regularization by dynamically calibrating the aggregation weights, CASC [17] utilizes layer weight to aggregate parameters.

3 Background

3.1 Federated Domain Generalization (FedDG).

As shown in Fig. 1(a), different from the conventional scenario, Federated Domain Generalization (FedDG) can only train several client models on local datasets. Given source domains $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_N\}$ where N denotes the number domains for training. Let (x, y) be a sample pair, x denotes a sample, and $y \in \{c_1, c_2, \dots, c_n\}$ is the corresponding one-hot label.

During local training, a baseline loss function \mathcal{L}_{ce} gauges the gap between the label and model prediction $f_k(x_k)$, where f_k represents client model for the k^{th} domain. Typically, this model entails as encoder for feature extraction and a simple fully-connected layer serving as the classifier. The standard training objective involves minimizing the cross-entropy loss \mathcal{L}_{ce}

$$\mathcal{L}_{ce} = -\frac{1}{M_k} \sum_1^{M_k} y_{k_i} \log(f(x_{k_i})) \quad (1)$$

where M_k denotes sample number of each client domain D_k and $(x_{k_i}, y_{k_i}) \in D_k$, $0 < i < M_k$.

Following local iterations, the global model $f(\cdot)$ undergoes updating via the aggregation of client model parameters. In the FedDG scenario, the global model confronts computer vision tasks within novel target domains, requiring adeptness in generalizing to fresh domains.

3.2 Diffusion Model

Diffusion model is a type of generative model that creates data by progressively adding noise to it and then learning to reverse the process to recover the data

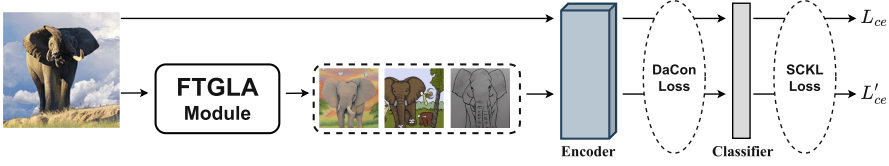


Fig. 2. Overall Framework of our method.

from the noise. The basic procedure of a diffusion model includes the forward diffusion process and the reverse denoising process.

The forward diffusion process is a Markov process that gradually adds noise to the data until it becomes nearly isotropic Gaussian noise. Given the original data x_0 and the noisy data at step t as x_t , the forward diffusion process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I) \quad (2)$$

where $\alpha_t \in (0, 1)$ is the scaling factor at time step t , typically set as a decreasing sequence. By combining multiple steps, the forward diffusion process can directly relate the original data x_0 , which can be described by:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (3)$$

where $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. The reverse denoising process involves training a model $p_\theta(x_{t-1}|x_t)$ to approximate the reverse of the forward process, where

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)) \quad (4)$$

where $\mu_\theta(x_t, t)$ and $\sum_\theta(x_t, t)$ are the mean and covariance parameterized by a neural network. During training, the diffusion model optimizes the parameters θ by minimizing the variational lower bound (VLB). A final a simplified loss that directly optimizes the reconstruction error at each step is often used:

$$L_{t-1} = \mathbb{E}_{x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right] \quad (5)$$

where $\epsilon_\theta(x_t, t)$ is the noise term predicted by the neural network.

During data generation, the process starts from a standard Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively applies the reverse denoising steps:

$$x_{t-1} \sim p_\theta(x_{t-1}|x_t) \quad (6)$$

4 Method

Figure 2 shows the overall framework of our method. Firstly, the input will undergo text-guided diffusion enhancement through the DACOA module, which will be input into the encoder together with the original image. After the encoder

extracts the feature representations, we introduce DaCon loss to shorten the similarity between features with same labels. After the features are input into a classifier to obtain the results, we design SCKL loss to ensure that the logits of the enhanced image are consistent with the original image. Also, the classification results of original and enhanced images are characterized by cross entropy loss. During the testing phase, the model only focuses on the classification results of the original image.

4.1 Diffusion-Aligned Coherent Augmentation (DACOA)

In FedDG scenario, a critical challenge is the inability to share information between domains due to privacy constraints. Previous methods attempted to exchange domain features between clients for augmentation, but the results had coherence issues. This exacerbates the difficulty of achieving generalization performance on unseen domains.

To address this issue, we propose the Diffusion-Aligned Coherent Augmentation (DACOA), shown in Algorithm 1. DACOA uses a pretrained latent diffusion model for text-guided data augmentation, which performs style transfer across different domains by text prompts. Firstly, the prompts follow the format of "a {domain} of {class}", such as "a cartoon of house". For a given domain, we randomly choose another domain of the same class to compose the prompt.

Algorithm 1 DACOA

Input: Domains $D = \{D_1, D_2, \dots, D_N\}$, CLIP text encoder T , Diffusion model M , Client models $f_i (i \in [1, N])$, Global model f

Output: Global model f

- 1: Initialize global model f parameters
 - 2: **for** total communication round **do**
 - 3: **for** each client model f_i **do**
 - 4: Distribute global model f parameters to client f_i
 - 5: **end for**
 - 6: **for** each client $i \in [1, N]$ **do**
 - 7: $D_{\text{other}} \leftarrow \{D_k \in D \mid k \neq D_i\}$
 - 8: **for** each sample $x \in D_i$ **do**
 - 9: $D_r \leftarrow \text{random_choice}(D_{\text{other}})$
 - 10: encode text as condition embedding $T_{\text{text}} \leftarrow T(\text{"a } D_r \text{ of } c_j\text{"})$
 - 11: encode image and noising $z \leftarrow M.\text{noising}(M.\text{encode}(x))$
 - 12: diffusion sample $\hat{z} \leftarrow M.\text{ddim}(z, T_{\text{text}})$
 - 13: decode back to image space $\hat{x} \leftarrow M.\text{decode}(\hat{z})$
 - 14: use (\hat{x}, x) to train client model f_i
 - 15: **end for**
 - 16: **end for**
 - 17: Update parameters of f by $f_i (i \in [1, N])$
 - 18: **end for**
-

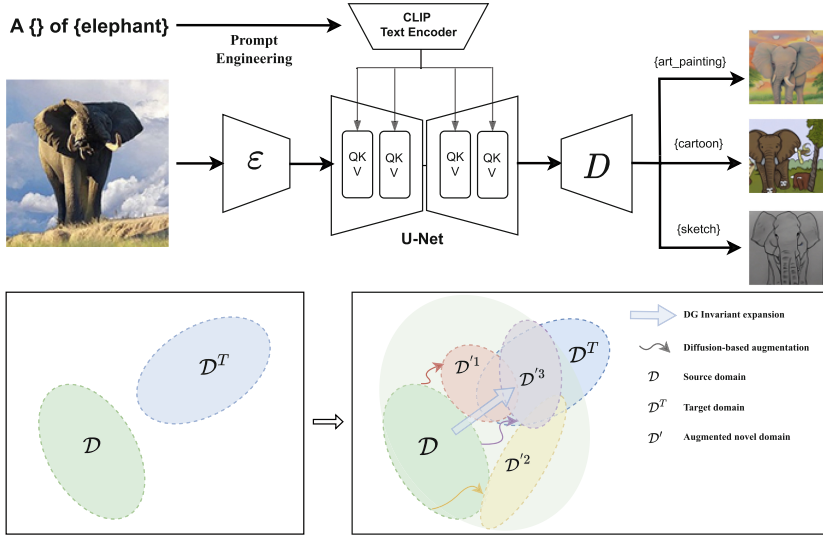


Fig. 3. Framework of DACOA and intuitive display of domain distribution. \mathcal{D} is the Source domain and \mathcal{D}^T is the target domain. DACOA uses text to guide the Diffusion model to generate different distributions \mathcal{D}' , thereby broadening the source and improving the coverage of the target domain.

After obtaining the prompt, we utilize CLIP’s text encoder to generate text embeddings, serving as conditional embeddings c during the denoising process of the latent diffusion model(LDM). Subsequently, following the setup of LDM, we encode and add noise to the input images, projecting them into the latent space. Finally, we employ conditional embeddings and image embeddings for the denoising operation within the diffusion process, written as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t, c) + \sigma_t\epsilon \tag{7}$$

Using text prompts for augmentation offers advantage of semantic consistency. The meanings of text prompt are fixed, ensuring that the generated images remain contextually appropriate for the specified class. Also, as shown in lower half of Fig. 3, the process of generating images by diffusion models involves random initial values, which expand the breadth of the source domain distribution. The randomness allows for diverse creation of the same image. Thus, even with a fixed text prompt, different initial values result in multiple variations of the augmented image, enhancing the diversity of the training data.

4.2 Domain aligning contrastive learning

In the context of privacy-preserving federated learning, data augmentation methods still face challenges in maintaining semantic consistency across different

domains. To address this limitation, incorporating supervised contrastive learning (SupCon) [8] into our training process can enhance the discriminative power of the model by leveraging label information to pull together samples from the same class and push samples apart from different classes. The SupCon loss allows the model to learn more distinct and robust feature representations, which is formulated as:

$$\mathcal{L}_{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \quad (8)$$

where $I \equiv \{1 \dots N\}$ denotes the set of indices of all features. $P(i)$ denotes the set of indices of all samples whose label is same as i . $A(i) \equiv \{1 \dots, i-1, i+1, \dots N\}$ denotes all indices but i .

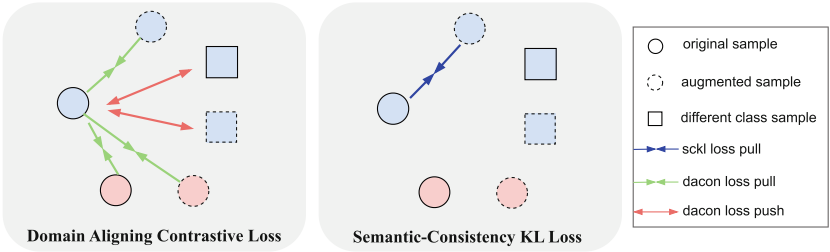


Fig. 4. Demonstration of DaCon and SCKL loss module, only the leftmost sample is used as an anchor in the figure. (I) In feature dimension learning, DaCon regularize the model to learn domain-invariant feature representations. (II) The enhanced image approximate the original classification result in the form of SCKL on the logits of the classification result, performing semantic constraints.

To further improve the feature extraction process, we bring another assumption that a robust feature extractor should embed augmented and original features adjacent to each other. We hence extend the concept of SupCon to Domain aligning contrastive(DaCon) loss, as shown in the left of Fig. 4. We minimize the distance between the original feature (anchor) and the augmented one in terms of the dot-product similarity. This strategy ensures both semantic alignment and class discriminability.

Our approach involves concatenating the original and augmented feature representations, resulting in a combined feature set. By creating class-based positive pairs and normalizing the feature vectors, we construct a similarity matrix to measure the relationships between the samples. The supervised contrastive loss is then computed by maximizing the similarity of samples within the same class while minimizing it for samples from different classes.

$$\mathcal{L}_{sup} = \sum_{i \in I \cup I'} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i) \cup A'(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \quad (9)$$

where I' and $A'(i)$ denotes augmented set of I and $A(i)$.

4.3 Semantic-Consistency KL Loss

Combining with DaCon mentioned above, we propose a Semantic-Consistency KL Loss (SCKL). Our approach aims to bridge the gap between different domains while adhering to privacy-preserving protocols. As shown in the right of Fig. 4, we achieve this by aligning the predicted logits of augmented data with those of the original data, thereby drawing them closer. Similar to knowledge distillation [5], we measure the KL divergence between the two sets of logits.

Unlike the classical distillation approach, where logits are typically softened to elevate the values of less likely categories, our method includes a sharpening operation on the original logits. This sharpening process ensures that the semantic information is preserved and highlighted during the learning process, which is crucial for maintaining the integrity of stylized predictions. By focusing on the semantic consistency between the original and augmented data, our method enhances the model’s ability to generalize across domains.

$$\mathcal{L}_{cons} = -\frac{1}{M_k} \sum_{i=1}^{M_k} f'_k(x_{k_i}) \log(f_k(x_{k_i}, \tau)) \quad (10)$$

where M_k is the number of samples (x_{k_i}, y_{k_i}) in D_k . Function $f_k(x, \tau)$ and $f'_k(x)$ denotes the predict results of original and augmented features. τ is set as 0.5 to perform square root of the predicted logits of f , denoting the sharpening approach. Finally, the complete loss can be written as:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}'_{ce} + \lambda_1 \mathcal{L}_{sup} + \lambda_2 \mathcal{L}_{cons} \quad (11)$$

5 Experiment

5.1 Implementation Detail

In this section, we choose three well-known domain generalization datasets and perform experiments on them:

- **PACS:** This dataset includes 9,991 images across seven classes, originating from four domains with distinct styles [9].
- **Office-Home:** Comprising 15,500 images, this dataset covers 65 classes across four different domains [13].
- **Digits-DG:** This dataset combines four traditional digit datasets [20].

For all benchmarks, we use a leave-one-domain-out evaluation strategy. In each round, one domain is treated as the unseen test domain while the remaining domains are used as the training sources. The training and validation splits within each source domain follow the configurations outlined in previous works [4, 19]. The entire unseen domain is utilized for testing purposes.

For the PACS and Office-Home datasets, we employ a ResNet-18 model pre-trained on ImageNet. For the Digits-DG dataset, we use a four-layer convolutional network as detailed in [19]. We use the pretrained, version 1.4 of stable

Table 1. Results on PACS datasets.

Paradigm	Method	PACS				
		Art	Cartoon	Photo	Sketch	Avg
Regular DG	Jigen [2]	79.40	75.30	96.00	71.40	80.50
	DDAIG [19]	84.20	78.10	95.30	74.70	83.10
	L2A-OT [20]	83.30	78.20	96.20	73.60	82.80
	MixStyle [21]	84.10	78.80	96.10	75.90	83.70
	EISNet [14]	80.00	76.00	93.70	80.90	82.60
	RISE [6]	85.10	81.80	96.00	78.40	85.30
Federated DG	FedDG [10]	83.94	79.27	96.23	73.30	83.19
	CASC [17]	82.00	76.40	95.20	81.60	83.80
	FADH [16]	83.80	77.20	94.40	84.40	85.00
	COPA [15]	83.30	79.80	94.60	82.50	85.10
	Baseline [18]	81.28	76.73	93.97	82.57	83.64
Our Method	FeSFD	84.26	81.97	96.35	84.81	86.85
	Improvement	↑ 2.98	↑ 5.24	↑ 2.38	↑ 2.24	↑ 3.21

Table 2. Results on Office-Home datasets.

Paradigm	Method	Office-Home				
		Artist	Clipart	Product	Real-world	Avg
Regular DG	Jigen [2]	53.00	47.50	71.50	72.80	61.20
	DDAIG [19]	59.20	52.30	74.60	76.00	65.50
	L2A-OT [20]	60.60	50.10	74.80	77.00	65.60
	MixStyle [21]	58.70	53.40	74.20	75.90	65.50
	EISNet [14]	56.80	53.30	72.30	73.50	64.00
	RISE [6]	59.10	52.90	75.10	76.40	65.90
Federated DG	FedDG [10]	60.70	45.82	71.51	73.05	62.77
	FedAvg [11]	58.20	51.60	73.10	73.80	64.20
	FADH [16]	59.90	55.80	73.50	74.90	66.00
	COPA [15]	59.40	55.10	74.80	75.00	66.10
	CCST [3]	59.05	50.06	72.97	71.67	63.56
	Baseline [18]	58.57	54.39	73.39	74.73	65.27
Our Method	DACOA	59.42	56.24	75.30	77.15	67.03
	Improvement	↑ 0.85	↑ 1.85	↑ 1.91	↑ 2.42	↑ 1.76

Table 3. Results on Digits-DG dataset. The best and second-best are bolded and underlined respectively.

Paradigm	Method	Digits-DG				
		MNIST	SVHN	SYN	MNIST-M	Avg
Regular DG	Jigen [2]	96.50	63.70	74.00	61.40	73.90
	DDAIG [19]	96.60	68.60	81.00	64.10	77.60
	L2A-OT [20]	96.70	68.60	83.20	63.90	78.10
	EISNet [14]	96.40	56.00	60.50	87.90	75.20
	MixStyle [21]	96.50	64.70	81.20	63.50	76.50
Federated DG	FedDG [10]	96.30	61.20	90.00	66.70	78.60
	FedAvg [11]	96.93	62.19	90.04	57.75	76.71
	FADH [16]	97.70	73.30	90.60	65.30	81.70
	COPA [15]	97.00	71.61	90.66	66.52	81.49
	CCST [3]	96.16	66.58	88.79	62.76	78.57
	Baseline [18]	96.52	62.80	90.42	59.16	77.23
Our Method	FeSFD	96.67	74.86	90.83	66.56	82.23
	Improvement	↑ 0.15	↑ 12.1	↑ 0.41	↑ 7.40	↑ 5.01

diffusion [12] without finetuning as our base LDM. Also, For Digits-DG dataset, we use “a new style of cls” as the prompt. To ensure fairness, we adhere to the protocol described in [4]. Following a strong baseline in [18], we dynamically calibrate client domains during model weight aggregation. The hyperparameters *strength*, λ_1 , λ_2 are set to 0.8, 5 and 0.3 respectively. All other hyperparameters and optimization settings are consistent with those specified in [18].

5.2 Domain Generalization Ability

We report the overall performance comparison of all methods in Table 1, 2 and 3, and our observations are summarized below:

- Compared with the strong baseline FedDG-GA [18], our proposed method improves the absolute value by 3.21%, 1.76% and 5.01% respectively, which effectively verifies the significant improvement of our proposed method in generalization ability.
- Compared to the best-performing methods among the three benchmarks, RISE, COPA, and FADH, our method outperforms them on average and has a significant improvement, indicating the superiority of our method.
- Our method has the most significant improvement on the Digits-DG dataset. We believe that the different setting of prompts have allowed DACOA to play a greater role, allowing local models to learn domain invariant features.

- Our method achieves the best performance in most single target domains. It can be seen that some methods surpass us on a single target domain, but have significant performance degradation on other target domains, so we have an advantage in overall generalization ability.

5.3 Ablation Study

Contributions of Different Components. As shown in Table 4, the baseline indicates FedAvg-GA [18] as our strong baseline. Line 2 shows that, by utilizing the DACOA augmentation strategy to expand client domains, the performance on unseen target domain is significantly improved. In Lines 3 and 4, the model improved by 0.89% and 1.13% respectively after adding DaCon and SCKL. It means that our method can make samples with the same label closer in terms of classification results or feature dimensions, thus learning semantic information. Line 5 shows that the combination of DaCon and SCKL can achieve optimal performance of the model, which demonstrates that these two modules we propose complement each other.

Table 4. Ablation study for Contributions of Different Components. Line one denotes the strong baseline we use.

\mathcal{L}_{ce}	DACOA	\mathcal{L}_{sup}	\mathcal{L}_{cons}	Acc
✓				83.64
✓	✓			85.21
✓	✓	✓		86.10
✓	✓		✓	86.34
✓	✓	✓	✓	86.85

Table 5. Ablation study for the strength DACOA executes after.

Strength	PACS				
	Art	Cartoon	Photo	Sketch	Avg
0.5	82.90	81.28	95.89	82.93	85.75
0.6	83.26	81.11	95.78	83.01	85.79
0.7	83.78	82.10	95.60	83.10	86.15
0.8	84.26	81.97	96.35	84.81	86.85
0.9	84.10	81.82	96.02	83.53	86.37

The influence of strength in LDM. Strength is an important parameter of image to image in LDM. Its meaning is: the number of times an image is added to noise and inference sampling divided by the number of inference sampling in the training process. Intuitively, the larger the strength setting, the closer the generated image will be to the description of conditional embedding, and it will also be more different from the original image. Table 5 shows the generalization ability effect of the model after selecting different strengths in the PACS dataset. It can be seen that the average result of generalization will gradually increase from 0.5 at the beginning, and the optimal result is when 0.8 is selected, but it starts to fall again when 0.9 is reached.

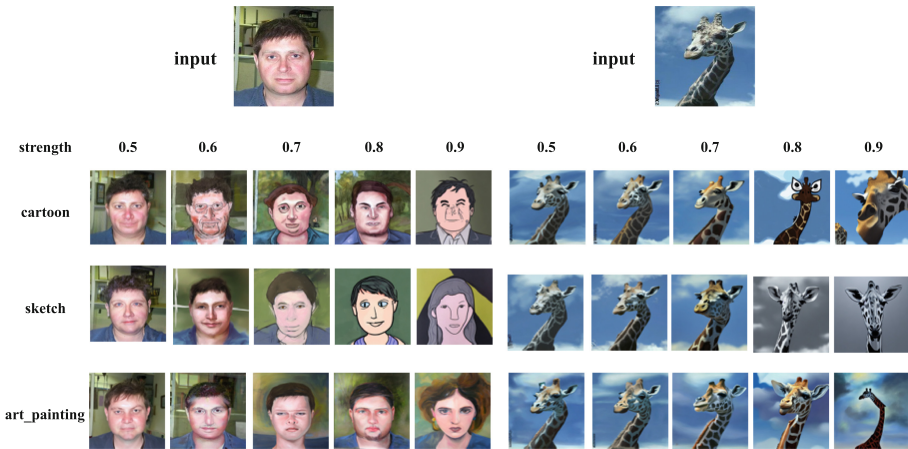


Fig. 5. Results of DACOA with different strength.

Figure 4 shows the generation result of the LDM under different strengths. It can be seen that when the strength is between 0.5 and 0.7, the generated image is often very close to the input. At 0.8, the image can have a good style transfer while maintaining original semantic. However, at 0.9, there will be significant semantic changes. For example, the left input of photo to sketch or photo to cartoon, the gender of the image has changed when strength is 0.9. In the right example, the output of the image have also been distorted or deformed. This indicates that the larger the augmentation amplitude is not the larger the better. If the strength is too large, the semantics of the image may even be changed, which affects the learning of models. In order to learn robust representation against domain shift, the model needs to achieve a balance between semantics and cross-style learning. Therefore, choosing a moderate strength of 0.8 will achieve the best generalization ability of the model.

5.4 Discussion

In this section, we analyze the shortcomings in the experiment and point out the parts that can be improved in the future. Specifically, for those instances where the model classifies incorrectly in both the original and augmented data, we observe the results of its style transfer. It can be found that stable diffusion still has problems as follows: (1) As shown in Figures 6.1 and 6.2, LDM omitted key data for classification during the generation process, resulting in significant unconscionable changes in the image. (2) In Figures 6.3, 6.4, and 6.5, LDM did not follow the guidance of the text prompt and generates incorrectly, which completely changes the semantics of the image. (3) In Figure 6.6, LDM’s inference was stopped early during a dynamic adjustment stage, resulting in distorted images.

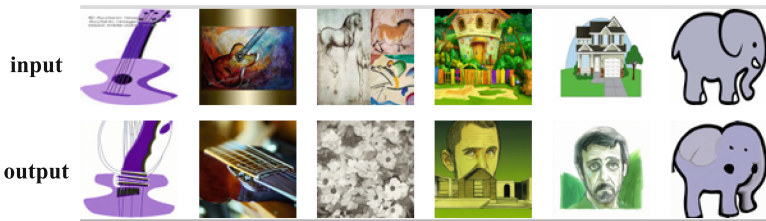


Fig. 6. DACOA augmentation results that are out of expectation.

In summary, relying solely on text prompts and adjusting strength parameters may still have limitations. We hope this work will encourage more research to address this issue and improve this novel framework. For example, fine-tune stable diffusion on each domain or use a higher quality stable diffusion model to control image quality during the generation phase. Additionally, a quality model can be introduced to exclude some negative samples that do not meet expectations. This will also be left for our future work.

6 Conclusion

In this paper, we propose DACOA, a diffusion-based augmentation method which can preserve image coherence, helping client models learn robust feature against domain shift. Additionally, clients can generate cross domain image using text prompt, so as not to violate privacy protocols. To conduct alignment on both the feature dimensions and prediction results, we introduce DaCon and SCKL loss, which improve the generalization ability of model. Adequate experiments and example diagrams demonstrate the superior effect of the model and provide guidance for how to use diffusion models in the DG field in the future.

Acknowledgments. This work is supported by Shenzhen Science and Technology Program (No. JCYJ20230807120800001)

References

1. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. *Advances in neural information processing systems* **19** (2006)
2. Carlucci, F.M., D’Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2229–2238 (2019)
3. Chen, J., Jiang, M., Dou, Q., Chen, Q.: Federated domain generalization for image recognition via cross-client style transfer. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 361–370 (2023)
4. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. *arXiv preprint [arXiv:2007.01434](https://arxiv.org/abs/2007.01434)* (2020)
5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint [arXiv:1503.02531](https://arxiv.org/abs/1503.02531)* (2015)
6. Huang, Z., Zhou, A., Ling, Z., Cai, M., Wang, H., Lee, Y.J.: A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11685–11695 (2023)
7. Jeon, S., Hong, K., Lee, P., Lee, J., Byun, H.: Feature stylization and domain-aware contrastive learning for domain generalization. In: *Proceedings of the 29th ACM International Conference on Multimedia*. pp. 22–31 (2021)
8. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. *Adv. Neural. Inf. Process. Syst.* **33**, 18661–18673 (2020)
9. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 5542–5550 (2017)
10. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1013–1023 (2021)
11. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: *Artificial intelligence and statistics*. pp. 1273–1282. PMLR (2017)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
13. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5018–5027 (2017)
14. Wang, S., Yu, L., Li, C., Fu, C.W., Heng, P.A.: Learning from extrinsic and intrinsic supervisions for domain generalization. In: *European Conference on Computer Vision*. pp. 159–176. Springer (2020)
15. Wu, G., Gong, S.: Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6484–6493 (2021)
16. Xu, Q., Zhang, R., Zhang, Y., Wu, Y.Y., Wang, Y.: Federated adversarial domain hallucination for privacy-preserving domain generalization. *IEEE Transactions on Multimedia* (2023)

17. Yuan, J., Ma, X., Chen, D., Wu, F., Lin, L., Kuang, K.: Collaborative semantic aggregation and calibration for federated domain generalization. *IEEE Transactions on Knowledge and Data Engineering* (2023)
18. Zhang, R., Xu, Q., Yao, J., Zhang, Y., Tian, Q., Wang, Y.: Federated domain generalization with generalization adjustment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3954–3963 (2023)
19. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34, pp. 13025–13032 (2020)
20. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: *Computer Vision—ECCV 2020: 16th European Conference, August 23–28, 2020, Proceedings*. pp. 561–578. Springer (2020)
21. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: *International Conference on Learning Representations* (2020)



Collaborative Domain Alignment for Multi-source Domain Adaptation

Yuanyuan Xu^{1,2(✉)}, Meina Kan^{1,2}, Zhilong Ji⁴, Jinfeng Bai⁴,
Shiguang Shan^{1,2,3}, and Xilin Chen^{1,2}

¹ Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

yuanyuan.xu@vip1.ict.ac.cn

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Peng Cheng Laboratory, Shenzhen, China

⁴ Tomorrow Advancing Life (TAL), Beijing, China

Abstract. In multi-source domain adaptation, the main challenge is effectively integrating information from various source domains and adapting it to the target domain. Existing methods either align feature distributions of each source domain with the target domain separately and fuse at the classifier level, or jointly align feature distributions of all domains. The former approach fragments shared information, while the latter sacrifices discriminative properties. To address this, we propose **Collaborative Domain Alignment (CoDA)**. CoDA utilizes an integrated feature encoder with domain attention masks to capture diverse shared information within a unified framework, thereby preserving both robustness and discriminability. Specifically, each source domain is elastically aligned with the target domain using a source-specific domain attention mask on the shared feature representation. Activated masks highlight features shared between individual source domains and the target domain, while overlapping masks highlight features shared by multiple source domains and the target domain. To optimize CoDA, we devise a domain-collaborative training strategy that includes domain-specific training loss, domain-consistency training loss, and pseudo-labeling loss. Extensive experiments on diverse datasets confirm the effectiveness and superiority of our approach.

Keywords: multi-source domain adaptation · domain adaptation · transfer learning

1 Introduction

Conventional machine learning assumes training and test data share the same distribution, leading to the expectation that a model performing well on training data will also perform well on test data. However, test data often differ from training data, as seen in robotic manipulation tasks where robots trained in simulations must operate in the real world [1, 23]. To tackle domain shift

and improve target domain performance, domain adaptation methods are used. While many focus on single-source domain adaptation (SDA), this study explores multi-source domain adaptation (MDA), which addresses scenarios with source data from diverse distributions.

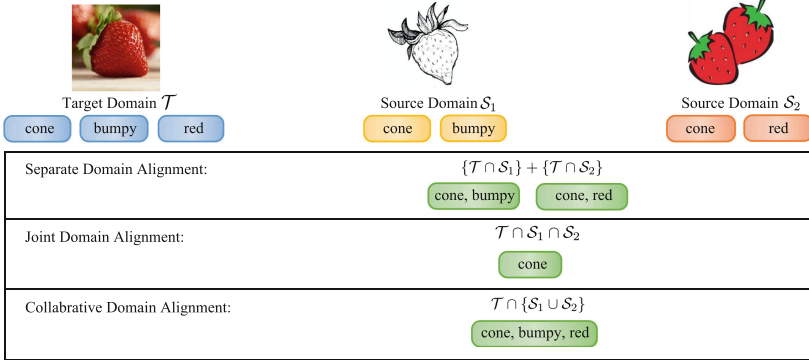


Fig. 1. A toy example compares different MDA methods. The goal is to improve the binary classification accuracy for the *strawberry* category in the target domain by utilizing labeled data from two source domains and unlabeled data from the target domain. Existing multi-source domain adaptation methods vary in how they leverage shared features to enhance the target task.

In MDA, the challenge is integrating information from multiple source domains and adapting it to the target domain. Existing methods fall into two categories: separate and joint domain alignment methods. Separate domain alignment methods [22, 29, 33] align each source domain independently with the target domain and then fuse different classification results, but they often overlook interactions between source domains [27], resulting in fragmented information. Joint domain alignment approaches [12, 13, 32] aim to learn features invariant to multiple domain shifts, but effectively filtering domain-specific information across various domains remains challenging and can lead to a loss of discriminative ability [30].

To address this issue, we propose **Collaborative Domain Alignment (CoDA)** for MDA. CoDA utilizes an integrated feature encoder with domain attention masks to capture diverse shared information within a unified framework, thereby preserving both robustness and discriminability. Fig. 1 illustrates CoDA’s advantage in classifying the *strawberry* category compared with other MDA methods. Unlike separate methods that learn pairwise features independently (e.g., $\{\mathcal{T} \cap \mathcal{S}_1\} + \{\mathcal{T} \cap \mathcal{S}_2\}$) or joint methods that focus only on features common to all domains (e.g., $\mathcal{T} \cap \mathcal{S}_1 \cap \mathcal{S}_2$), CoDA integrates attributes shared by all domains and those shared by partial domains, similar to $\mathcal{T} \cap \{\mathcal{S}_1 \cup \mathcal{S}_2\}$.

To learn such an integrated feature encoder, each source domain aligns flexibly with the target domain using source-specific domain attention masks applied

to the shared feature representation. These masks highlight features common to individual source domains and the target domain, while overlaps reveal features shared across multiple source domains and the target domain. We optimize the encoder with a domain-collaborative training strategy, which includes three types of losses: a domain-specific training loss for extracting shared features; a domain-consistency training loss for domain collaboration; and a pseudo-labeling loss for improving feature discriminability.

2 Related Work

Single-source domain adaptation addresses the domain shift between a single source and the target domain, forming the foundation for multi-source domain adaptation. CoDA utilizes domain attention masks, which are closely related to attention mechanisms. In this section, we will introduce these related methods.

Single-Source Domain Adaptation. Deep SDA methods primarily minimize domain shift by mapping source and target data into a shared latent feature space, categorized into discrepancy-based [15, 17] and adversarial-based approaches [6, 9]. Discrepancy-based methods reduce domain discrepancy by aligning first or second order data statistics. For example, DAN (deep adaptation network) [15] minimizes the maximum mean discrepancy in the reproducing kernel Hilbert space between two feature distributions. Adversarial-based methods reduce the domain gap by extracting domain-invariant features through adversarial learning. For example, DANN (domain-adversarial training of neural networks) [6] learns invariant features through gradient inversion. Recent SDA works [3, 10, 16] further focus on improving feature discriminability when learning domain-invariant features. However, SDA methods are not effective when dealing with multiple source domains as information fusion is not considered.

Multi-Source Domain Adaptation. Deep MDA methods can be broadly classified into separate and joint domain alignment methods. Separate domain alignment methods [21, 22, 29, 33] align the distribution of the target domain with each source domain independently, and the final result is obtained by fusing different classification predictions. For example, DCTN (deep cocktail network) [29] deploys multi-way adversarial learning to align the target domain with each source domain. Joint domain alignment methods [12, 13, 32] jointly align feature distributions of all domains, aiming to extract features that are agnostic across all domains. For example, DRT (dynamic residual transfer) [13] uses a dynamic network instead of a static one to align the target domain with all source domains, to better handle conflicts across multiple domains. Several studies [4, 5, 35] have introduced attention mechanisms into MDA. For example, DAC-Net (domain attention consistency network) [5] uses a feature channel attention module to emphasize transferable attributes for the target domain. However, our proposed CoDA stands apart by using attention mechanisms to preserve diverse features.

Attention Mechanism. Attention mechanisms, inspired by human information processing, are categorized into channel, spatial, temporal, and branch types [20]. Our work focuses on channel attention, which adjusts channel weights adaptively. Channel attention has been studied across various fields [2, 8, 28]. Our CoDA method is particularly related to DMG (domain-specific masks for generalization) [2]. Unlike DMG, which uses activation masks learned independently of the target domain, CoDA learns masks specifically related to the target domain and adds a domain consistency loss to enhance domain collaboration.

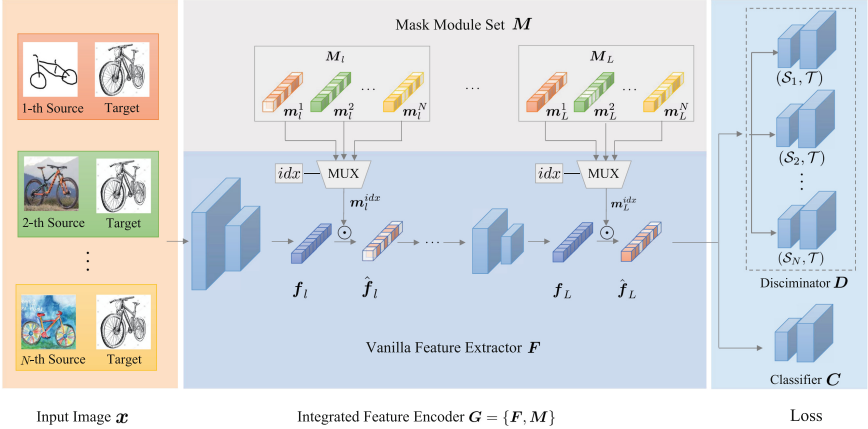


Fig. 2. An overview of CoDA in the training phase. CoDA consists of three components: an integrated feature encoder \mathbf{G} to extract diverse features, a multi-source domain discriminator \mathbf{D} to ensure the extracted features are domain-invariant, and a common category classifier \mathbf{C} to make the extracted features discriminative for classification. \mathbf{G} includes a vanilla feature extractor \mathbf{F} and a mask module set \mathbf{M} . During training, the network takes a single image \mathbf{x} and its corresponding mask index idx to extract features shared between the idx -th source domain and the target domain. For a specific layer l , the output feature \hat{f}_l of \mathbf{G} at that layer is obtained by channel-wisely multiplying the feature f_l extracted by \mathbf{F} with m_i^{idx} selected by the multiplexer (MUX).

3 Problem Setup

Specifically, there are N labeled source domains $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_N\}$ and one unlabeled target domain \mathcal{T} . The samples of N source domains are drawn from N different distributions, and $\mathcal{S}_i = \{(\mathbf{x}_k^i, \mathbf{y}_k^i)\}_{k=1}^{|\mathcal{S}_i|}$ stands for samples of the i -th source domain, where \mathbf{x}_k^i is a raw image and \mathbf{y}_k^i is the corresponding category label. For the target domain \mathcal{T} , only unlabeled images $\{\mathbf{x}_k^T\}_{k=1}^{|\mathcal{T}|}$ are accessible. The classification tasks of all domains are the same, and the goal of multi-source domain adaptation is to build a model that performs well on the target domain using both labeled source samples and unlabeled target samples.

4 Collaborative Domain Alignment

While each source domain shares some beneficial information with other domains (e.g., *cone* attribute depicted in Fig. 1), it also possesses unique information (e.g., *bumpy* or *red*) that can be advantageous for the target domain. Integrating all beneficial information is challenging due to: (1) *it is hard to mitigate various domain shifts while maintaining feature discriminability*, and (2) *it is unclear how to identify which information is shared among all domains and which is only present in partial domains*. CoDA addresses these challenges with an integrated feature encoder \mathbf{G} , which comprises a vanilla encoder \mathbf{F} and source-specific domain attention masks \mathbf{M} . These binary masks enable domain-specific alignment between the target domain and each source domain, resulting in the preservation of discriminative features that contain diverse beneficial information. Additionally, they facilitate domain collaboration and automatically determine which information is shared by all domains and which is shared by partial domains. Besides \mathbf{G} , the overall framework also includes a multi-source domain discriminator $\mathbf{D} = \{\mathbf{D}_i\}_{i=1}^N$ and a common category classifier \mathbf{C} , as depicted in Fig. 2. Next, we introduce the integrated feature encoder \mathbf{G} and the training scheme using domain-collaborative training.

4.1 Integrated Feature Encoder \mathbf{G}

The integrated feature encoder \mathbf{G} extracts features shared across all domains and across partial domains. It includes a vanilla feature extractor \mathbf{F} , composed of convolutional and fully connected layers, and a set of mask modules $\mathbf{M} = \{\mathbf{M}_1, \dots, \mathbf{M}_L\}$, where \mathbf{M}_l is the mask module at layer l , containing N source-specific domain attention masks $\mathbf{M}_l = \{\mathbf{m}_l^1, \dots, \mathbf{m}_l^N\}$, with \mathbf{m}_l^i for the i -th source domain. The masks are designed similarly across layers, so next we omit the subscript for brevity to illustrate the design, i.e. $\{\mathbf{m}^1, \dots, \mathbf{m}^N\}$.

During training, the inputs to the network consist of an image \mathbf{x} and its corresponding mask index idx . For an image from the i -th source domain (denoted as \mathbf{x}^i), idx is set to i . The source-specific domain attention mask \mathbf{m}^i is then selected by a multiplexer (MUX). The shared feature between the i -th source domain and the target domain at layer l is then computed as follows:

$$\hat{\mathbf{f}}^i = \mathbf{G}_l(\mathbf{x}^i; \mathbf{m}^i) = \mathbf{f}^i \odot \mathbf{m}^i. \quad (1)$$

Here, \mathbf{f}^i is the feature of \mathbf{x}^i extracted by the vanilla feature extractor \mathbf{F} , and $\mathbf{m}^i \in \{0, 1\}^K$ is the selected mask, where K is the number of feature channels or neurons of \mathbf{f}^i . The mask \mathbf{m}^i is multiplied channel-wisely with the feature map \mathbf{f}^i . This process can be repeated across all layers $\{1, 2, \dots, L\}$ until the final masked feature $\hat{\mathbf{f}}_L^i$ is obtained, or it can be applied to specific layers.

For a target domain image \mathbf{x}^T , the mask index idx varies depending on the desired shared features. If the shared features between the i -th source domain and the target domain are needed, idx is set to i , and the attention mask \mathbf{m}^i is chosen by the multiplexer (MUX) to calculate the shared feature:

$$\hat{\mathbf{f}}^{T|i} = \mathbf{G}_l(\mathbf{x}^T; \mathbf{m}^i) = \mathbf{f}^T \odot \mathbf{m}^i. \quad (2)$$

Here, the binary masks are learned by a thresholding function as Piggyback [19]. Specifically, each mask vector \mathbf{m}^i is associated with a learnable real-valued vector $\tilde{\mathbf{m}}^i$. In the forward pass, \mathbf{m}^i is obtained by setting a threshold on $\tilde{\mathbf{m}}^i$:

$$\mathbf{m}^i(j) = \begin{cases} 1, & \text{if } \tilde{\mathbf{m}}^i(j) \geq \tau \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where j is the j -th dimension and τ is a threshold. Since the thresholding function is non-differentiable, the gradients of the real-valued mask vector $\tilde{\mathbf{m}}^i$ are approximated by the gradients of the thresholded mask vector \mathbf{m}^i .

Naturally, masks for different source domains often overlap. Since the feature extractor \mathbf{F} is unified for all domains, overlapping masks reveal feature channels shared by multiple source domains and the target domain. To quantify this, we sum the source-specific domain attention masks into \mathbf{m}' :

$$\mathbf{m}' = \sum_{i=1}^N \mathbf{m}^i. \quad (4)$$

Here, $\mathbf{m}'(j)$ indicates how many source domains activate the j -th channel or node in \mathbf{F} . Specifically, $\mathbf{m}'(j) = N$ means the feature channels are shared by all domains (e.g., the *cone* attribute in Fig. 1), $\mathbf{m}'(j) = 1$ means the channels are shared by the target domain and one source domain, and $1 < \mathbf{m}'(j) < N$ means they are shared by the target domain and some source domains. Thus, \mathbf{G} can preserve features shared by all or some source domains and the target domain, capturing diverse shared features.

4.2 Domain-collaborative Training

The integrated feature encoder \mathbf{G} preserves diverse shared information across domains. It is trained using a domain-collaborative training strategy with three losses: a domain-specific loss for extracting shared information between the target and each source domain, leveraging mask overlaps for collaborative optimization; a domain-consistency loss to enhance source domain collaboration; and a pseudo-labeling loss to improve feature discriminability for the target domain.

Domain-specific Training Loss. The domain-specific training loss aims to optimize the model for extracting shared features between the target and each source domain using source-specific domain attention masks. For clarity, consider the alignment between the i -th source domain and the target domain. Given an instance \mathbf{x} from either \mathcal{S}_i or \mathcal{T} , the shared features are computed as $\hat{\mathbf{f}}_L^i$ and $\hat{\mathbf{f}}_L^{T|i}$ (Eq.(1) and (2)). These features should be domain-invariant and discriminative for classification. To ensure domain invariance, an adversarial loss is applied:

$$\begin{aligned} \mathcal{L}_i^{adv}(\mathbf{D}_i, \mathbf{G}) = & -\mathbb{E}_{\mathbf{x} \sim \mathcal{S}_i} \log \left[\mathbf{D}_i \left(\hat{\mathbf{f}}_L^i, \mathbf{p}^i \right) \right] \\ & -\mathbb{E}_{\mathbf{x} \sim \mathcal{T}} \log \left[1 - \mathbf{D}_i \left(\hat{\mathbf{f}}_L^{T|i}, \mathbf{p}^T \right) \right]. \end{aligned} \quad (5)$$

Here, \mathbf{p}^i and \mathbf{p}^T are the softmax outputs of \mathbf{C} for the masked features $\hat{\mathbf{f}}_L^i$ and $\hat{\mathbf{f}}_L^{T|i}$ extracted by \mathbf{G} . Both features and classification results are fed into \mathbf{D}_i for category-level domain alignment, as in CDAN [16]. The loss in Eq. (5) is minimized for \mathbf{D}_i and maximized for \mathbf{G} , training \mathbf{G} to generate domain-invariant features that confuse \mathbf{D}_i .

To enhance the discriminability of shared features, the classification loss is computed for the source feature $\hat{\mathbf{f}}_L^i$ from \mathbf{G} using the cross-entropy between the classifier \mathbf{C} 's predicted probabilities \mathbf{p}^i and the true label \mathbf{y} :

$$\mathcal{L}_i^{cls}(\mathbf{C}, \mathbf{G}) = \mathbb{E}_{(x, \mathbf{y}) \sim \mathcal{S}_i} L_{ce}(\mathbf{p}^i, \mathbf{y}). \tag{6}$$

The overall domain-specific training loss is formulated as follows:

$$\begin{aligned} \mathcal{L}^{spf}(\mathbf{C}, \mathbf{D}, \mathbf{G}) &= \frac{1}{N} \sum_{i=1}^N \mathcal{L}_i^{spf}(\mathbf{C}, \mathbf{D}, \mathbf{G}) \\ &= \frac{1}{N} \sum_{i=1}^N [-\mathcal{L}_i^{adv}(\mathbf{D}_i, \mathbf{G}) + \mathcal{L}_i^{cls}(\mathbf{C}, \mathbf{G})], \end{aligned} \tag{7}$$

where the domain-specific training loss \mathcal{L}^{spf} is calculated by averaging the individual losses \mathcal{L}_i^{spf} (for $i \in \{1, 2, \dots, N\}$) across all source-target pairs.

When a feature channel is activated by multiple source-specific domain attention masks, it indicates the feature channel is shared among the target and several source domains. Consequently, the model parameters for this feature channel are updated by gradients from all involved domains, enabling collaborative optimization of the feature encoder \mathbf{F} . Fig.3 shows this: for example, the fourth node, activated by three source domains and the target domain, is optimized using losses from those domains. This collaborative updating is a kind of implicit domain collaboration.

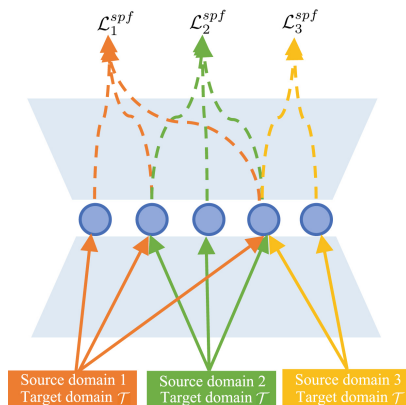


Fig. 3. An illustration of domain-specific training loss and collaborative optimization.

Domain-consistency Training Loss. While overlapped domain attention masks help the domain-specific training loss preserve diverse shared information and encourage source domain collaboration, they don't ensure consistent feature values for shared features. For example, in Fig.3, a target sample processed through the fourth node with different source-specific masks might yield varying feature values. This variability is problematic, as the final target features should be consistent at shared positions for accurate classification.

To address this, a domain-consistency training loss is applied to the target domain's final layer. This loss ensures that target features extracted with different masks remain consistent at overlapping positions, promoting unanimous domain collaboration. The domain-consistency training loss is defined as follows:

$$\mathcal{L}^{con}(\mathbf{G}) = \frac{1}{P} \sum_{(i,j)} \left\| (\hat{\mathbf{f}}_L^{T|i} - \hat{\mathbf{f}}_L^{T|j}) \odot (\mathbf{m}_L^i \odot \mathbf{m}_L^j) \right\|_1, \quad (8)$$

where $\hat{\mathbf{f}}_L^{T|i}$ and $\hat{\mathbf{f}}_L^{T|j}$ indicate the features of a target sample when extracted with domain attention masks from the i -th and j -th source domain, and P is a constant that is calculated by multiplying the number of source pairs by the number of feature channels to normalize the domain-consistency training loss.

Pseudo-labeling Loss. The domain-specific and domain-consistency training losses help capture diverse features in the target domain. To enhance feature discriminability, pseudo-labels are used to guide learning. Pseudo-labels with confidence above 0.9 are generated from target domain features using K-means clustering. Confidence is derived by converting distances to classification probabilities, similar to MLAN [31]. These pseudo-labels are then used to compute the classification loss for the target domain:

$$\mathcal{L}^{pse}(\mathbf{C}, \mathbf{G}) = \mathbb{E}_{(\mathbf{x}, \hat{\mathbf{y}}) \sim \mathcal{T}_p} L_{ce}(\mathbf{p}^T, \hat{\mathbf{y}}), \quad (9)$$

where \mathcal{T}_p is the set of target samples with pseudo-labels, and $(\mathbf{x}, \hat{\mathbf{y}})$ represents a target image and its pseudo label from \mathcal{T}_p .

Overall, by combining the domain-specific training loss, the domain-consistency loss, and the pseudo-labeling loss in Eq.(7), Eq.(8), and Eq.(9), the training loss of the whole proposed CoDA method is formulated as follows:

$$\min_{\mathbf{C}, \mathbf{G}=\{\mathbf{F}, \mathbf{M}\}} \mathcal{L}^{spf}(\mathbf{C}, \mathbf{D}, \mathbf{G}) + \lambda \mathcal{L}^{con}(\mathbf{G}) + \mathcal{L}^{pse}(\mathbf{C}, \mathbf{G}), \quad (10)$$

$$\min_{\mathbf{D}} -\mathcal{L}^{spf}(\mathbf{C}, \mathbf{D}, \mathbf{G}), \quad (11)$$

where λ is the weight hyperparameter of the domain-consistency training loss.

4.3 Testing

During training, the features of the target domain are respectively extracted under different domain masks to determine shared features with each source

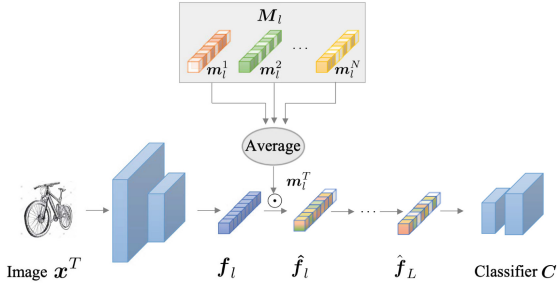


Fig. 4. CoDA in the testing phase. Given a sample x^T from the target domain during testing, the vanilla feature extractor F extracts the feature f_l at layer l . A soft attention mask m_l^T is applied to f_l , which is obtained by averaging all source-specific domain attention masks at that layer.

domain. During testing, all shared features should be utilized for a more accurate classification of the target domain. However, the integrated feature encoder G has only been trained on partially activated target features and has not been exposed to all activated feature channels. For instance, during training, G_{l+1} only uses $f_l^T \odot m_l^i$ ($i \in \{1, 2, \dots, N\}$) but not f_l^T directly, potentially degrading accuracy at test time. To address this, a soft-scaling scheme inspired by DropOut [2, 26] is used, where the test mask m^T is the average of all source-specific domain attention masks:

$$m^T = \frac{1}{N} \sum_{i=1}^N m^i. \quad (12)$$

This scheme is applied on all layers that have domain attention masks. By ensuring that the expectations for the output remain consistent during both training and testing, the mismatch problem can be alleviated.

5 Experiments

In this section, we evaluated the effectiveness of CoDA by conducting experiments on three commonly used datasets: DomainNet, PACS, and Office-31.

5.1 Setup

Datasets. DomainNet [22] is a large-scale dataset with about 0.6 million images across 345 categories from 6 domains: Clipart (clp), Infograph (inf), Painting (pnt), Quickdraw (qdr), Real (rel), and Sketch (skt). PACS [11] contains 9,991 images in 7 categories from 4 domains: Photo (P), Art Painting (A), Cartoon (C), and Sketch (S). Office-31 [25] includes 4,652 images of 31 object categories from 3 domains: Amazon (A), Webcam (W), and Dslr (D).

Implementation Details. To ensure a fair comparison, ResNet-18, ResNet-50, and ResNet-101 [7] are used as the backbone models for the PACS, Office-31,

and DomainNet datasets, respectively, which is consistent with previous studies [13, 27, 29]. In terms of hyperparameters, the weight λ of the domain-consistency training loss in Eq.(10) is set to 0.5 for all three datasets. By default, the mask module set $\mathbf{M} = \{\mathbf{M}_1, \dots, \mathbf{M}_L\}$ are applied to 4 layers of the vanilla feature encoder \mathbf{F} , i.e. the final layers of conv3_x, conv4_x, and conv5_x in the ResNet architecture and the bottleneck layer. Note that the additional computational overhead of CoDA on the feature extractor involves only a few element-wise feature multiplication operations (Eq.(1) and (2)) and is therefore negligible. The associated real-valued weight of each mask vector in Eq.(3) is initialized with 1e-2, and the binary threshold τ is set to 5e-3 following Piggyback [19]. Regarding optimization details, the proposed CoDA is implemented using Pytorch. The Adam optimizer is used, with a small learning rate of 1e-5 for pre-trained parameters and a faster learning rate (10 \times) for the rest. The batch size is set to 32 for all experiments. For smaller datasets like PACS and Office-31, the model is trained for 30 epochs, and the training time is less than 1 hour for a specific transfer task. For the larger DomainNet dataset, the model is trained for 10 epochs, and the training time is approximately 16 hours per task.

Table 1. Classification accuracy (%) on DomainNet dataset

Methods	\rightarrow clp	\rightarrow inf	\rightarrow pnt	\rightarrow qdr	\rightarrow rel	\rightarrow skt	Avg
MDAN [32]	52.4	21.3	46.9	8.6	54.9	46.5	38.4
M ³ SDA [22]	58.6	26.0	52.3	6.3	62.7	49.5	42.6
MDDA [33]	59.4	23.8	53.2	12.5	61.8	48.6	43.2
ML-MSDA [14]	61.4	26.2	51.9	19.1	57.0	50.3	44.3
T-SVDNet[21]	66.1	25.0	54.3	16.5	65.4	54.6	47.0
LtC-MSDA [27]	63.1	28.7	56.1	16.3	66.1	53.8	47.4
DCTN [29]	69.6	27.5	57.3	17.8	72.5	55.3	49.8
DRT[13]	69.7	31.0	59.5	9.9	68.4	59.4	49.7
DRT+ST[13]	71.0	31.6	61.0	12.3	71.4	60.7	51.3
PTMDA[24]	66.0	28.5	58.4	13.0	63.0	54.1	47.2
ADNT[30]	69.0	28.3	60.5	16.3	68.7	63.5	51.0
DAC-Net [5]	72.5	27.6	57.8	23.0	66.7	59.5	51.2
MLAN [31]	71.4	29.3	59.5	28.4	73.9	58.7	53.5
CoDA (ours)	71.3	28.9	60.1	29.6	73.1	60.8	54.0

5.2 Comparisons to the State-of-the-art

In this section, we first compare CoDA with existing multi-source domain adaptation methods.

Results on DomainNet Dataset. CoDA achieves an average accuracy of 54.0% on the DomainNet dataset, surpassing all other state-of-the-art methods.

Table 2. Classification accuracy (%) on PACS dataset

Methods	→ A	→ C	→ S	→ P	Avg
MDAN [32]	83.5	82.3	72.4	92.9	82.8
MDDA [33]	86.7	86.2	77.6	93.9	86.1
DCTN [29]	84.7	86.7	71.8	95.6	84.7
M ³ SDA [22]	84.2	85.7	74.6	94.5	84.7
T-SVDNet [21]	90.4	90.6	85.5	98.5	91.3
DAC-Net [5]	91.4	91.4	85.0	97.9	91.4
CoDA (ours)	92.4	91.0	87.0	98.2	92.1

Table 3. Classification accuracy (%) on Office-31 dataset

Methods	→ D	→ W	→ A	Avg
DCTN [29]	99.3	98.2	64.2	87.2
M ³ SDA [22]	99.3	98.0	67.2	88.2
LtC-MSDA [27]	99.4	97.7	68.6	88.6
M ³ SDA- β [22]	99.6	99.3	69.4	89.5
MFSAN [34]	99.5	98.5	72.7	90.2
ADNT [30]	100	99.6	74.4	91.3
MLAN [31]	99.6	98.8	75.7	91.4
CoDA (ours)	99.5	98.9	74.4	90.9

Although no single approach excels in every transfer task, CoDA performs consistently well across all tasks. Notably, it shows strong performance on ‘qdr’, likely because it is distinct from other domains and benefits from relaxed alignment and K-means based pseudo labels.

Compared to other methods, CoDA outperforms all separate domain alignment methods, surpassing the best one, DCTN [29], by 4.2%. This is due to CoDA’s effective integration of diverse information within a unified model, allowing parameters to be updated with samples from multiple domains, resulting in more accurate representations. Additionally, CoDA surpasses joint domain alignment methods, including the best one, DRT+ST [13], by 2.7%, due to its superior ability to preserve diverse shared features.

While CoDA only has a slight 0.5% lead over the mutual learning based method MLAN [31], it is significantly simpler, employing a unified framework with a single feature extractor, unlike MLAN, which uses multiple feature extractors. Overall, these results highlight the benefits of collaborative learning among multiple source domains.

Results on PACS Dataset. In Tab.2, CoDA achieves the highest average accuracy of 92.1% on the PACS dataset, outperforming existing MDA methods. It performs best on ‘→ A’ and ‘→ S’ tasks and second-best on ‘→ C’ and ‘→ P’.

These results highlight the superiority of our approach. CoDA’s clear advantage on both DomainNet and PACS datasets is due to the distinct stylistic variations across domains, allowing it to leverage diverse shared information for efficient collaboration among source domains.

Results on Office-31 Dataset. In Tab. 3, CoDA achieves an average accuracy of 90.9% on the Office-31 dataset, comparable to the state-of-the-art method MLAN but slightly behind. This is likely because Office-31 has only three domains with relatively small gaps, offering less diverse shared information than other datasets, making CoDA’s advantage less evident.

Table 4. Ablation Study on PACS dataset

Method	\mathcal{L}^{spf}	\mathcal{L}^{con}	\mathcal{L}^{pse}	$\rightarrow A$	$\rightarrow C$	$\rightarrow S$	$\rightarrow P$	Avg
Baseline			✓	91.2	88.7	82.9	97.8	90.1
CoDA (w/o \mathcal{L}^{con})	✓		✓	91.3	87.7	85.0	97.9	90.5
CoDA (full version)	✓	✓	✓	92.4	91.0	87.0	98.2	92.1

5.3 Analysis

Ablation Study. In CoDA, the collaborative training strategy includes three losses: domain-specific training loss \mathcal{L}^{spf} , domain-consistent training loss \mathcal{L}^{con} , and pseudo-labeling loss \mathcal{L}^{pse} . The ablation study in Tab. 4 evaluates their efficacy by gradually adding each loss based on \mathcal{L}^{pse} . For the baseline method in the first row of Tab.4, \mathcal{L}^{spf} is removed by setting each mask dimension $m^i (i \in 1, \dots, N)$ to 1 in Eq.(3), causing \mathcal{L}^{spf} in Eq. (7) to degrade to joint domain alignment. The baseline model achieved an average accuracy of 90.1%. The second row introduces \mathcal{L}^{spf} , achieving a slight improvement of 0.4%, demonstrating its benefit to the integrated feature encoder. However, the method in the second row shows a 1.0% drop when transferring to task ‘ $\rightarrow C$ ’, indicating inconsistent cooperation between source domains. The last row presents the full version of CoDA, adding the domain-consistency training loss \mathcal{L}^{con} , which improves performance by 1.6% over the second row and 2.0% over the first row, highlighting the benefits of incorporating \mathcal{L}^{con} for better model performance. Note that \mathcal{L}^{con} must be added only after incorporating \mathcal{L}^{spf} since \mathcal{L}^{con} depends on \mathcal{L}^{spf} . Without \mathcal{L}^{spf} , \mathcal{L}^{con} would be zero and unnecessary. Therefore, although the improvement from \mathcal{L}^{spf} is modest, it is essential because \mathcal{L}^{con} relies on it.

To visually understand the influence of the two losses, t-SNE (t-distributed stochastic neighbor embedding) [18] is used to visualize the feature distributions of different methods. Fig. 5(a) illustrates the features of the baseline model, which correspond to the first row of results in Tab.4. In this figure, features from different domains but belonging to the same category are mixed together, as F only utilizes domain-invariant features among all domains. In contrast, Fig.

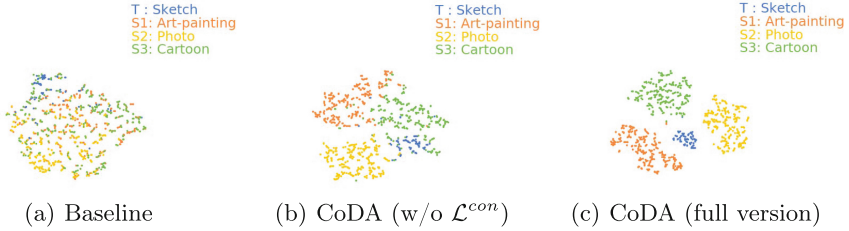


Fig. 5. Visualization of features from the ‘horse’ category for the ‘ \rightarrow S (Sketch)’ transferring task on the PACS dataset using t-SNE. The features before the classifier are used to compute t-SNE. Different colors stand for different domains.

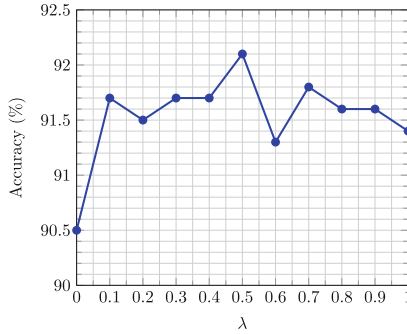


Fig. 6. Sensitivity analysis experiments of λ . Results are reported on PACS dataset.

5(b) shows the features generated by the integrated feature encoder \mathbf{G} using only the implicit domain collaboration loss, corresponding to the second row of results in Tab.4. These features exhibit diverse intra-class variance due to domain diversity. Finally, Fig. 5(c) presents the features obtained from the full version of CoDA, which corresponds to the last row of results in Tab.4. Here, features from different domains are pieced together with slight separation, while features from the same domain are grouped together. This demonstrates that features shared by partial domains are more effectively preserved by \mathbf{G} , leading to a better feature representation with the explicit domain collaboration restriction, i.e. the domain consistency training loss.

Sensitivity of Hyper-parameter λ . Fig. 6 shows CoDA’s sensitivity to the hyper-parameter λ in Eq. (10), which weights the domain-consistency training loss. A λ value of zero means no domain consistency restriction. Increasing λ from 0 to 0.1 improves target domain performance, demonstrating the loss’s effectiveness. However, accuracy variation is minimal within the range of 0.1 to 1.0, indicating CoDA’s robustness to changes in λ within this interval.

Effect of the Number of Mask Modules. Tab.5 shows the impact of adding mask modules on model performance. Mask modules, which can be added to any layer of the vanilla feature extractor \mathbf{F} , improve domain cooperation. In

Table 5. Effect of the number of mask modules

Number	→ A	→ C	→ P	→ S	Avg
1	90.9	88.7	84.4	97.7	90.4
2	91.4	91.5	81.7	98.1	90.7
3	91.9	90.4	83.8	98.3	91.1
4	92.4	91.0	87.0	98.2	92.1
5	92.7	90.0	85.4	98.2	91.6
6	91.4	91.4	86.2	98.0	91.7

ResNet architectures [7], they can be installed in up to 6 layers: conv1_x through conv5_x and an additional bottleneck layer. Results show that performance increases with up to 4 mask modules but slightly declines with more. This indicates that while additional mask modules enhance cooperation, their benefits diminish beyond a certain point, probably because features in lower layers are often already diverse enough.

6 Conclusion and Future Work

In conclusion, we introduce collaborative domain alignment (CoDA), a method that integrates features shared across all domains and those shared by subsets of domains. Our results show that preserving diverse shared information improves the performance of the target domain. CoDA outperforms state-of-the-art methods on several benchmark datasets, making it a promising solution for multi-source domain adaptation challenges. Meanwhile, the method has some limitations, such as involving a relatively large number of hyperparameters, which can make the tuning process more complex. Future work will focus on simplifying this process to improve usability.

Acknowledgements. This work was partially supported by the Natural Science Foundation of China (Nos. U2336213 and 62122074) and the Innovation Funding of ICT CAS (Nos. E000000 and E461010).

References

1. Arndt, K., Hazara, M., Ghadirzadeh, A., Kyrki, V.: Meta reinforcement learning for sim-to-real domain adaptation. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 2725–2731 (2020)
2. Chattopadhyay, P., Balaji, Y., Hoffman, J.: Learning to balance specificity and invariance for in and out of domain generalization. In: European Conference on Computer Vision (ECCV). pp. 301–318 (2020)
3. Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: International Conference on Machine Learning (ICML). pp. 1081–1090 (2019)

4. Cui, X., Bollegala, D.: Multi-source attention for unsupervised domain adaptation. arXiv preprint [arXiv:2004.06608](https://arxiv.org/abs/2004.06608) (2020)
5. Deng, Z., Zhou, K., Yang, Y., Xiang, T.: Domain attention consistency for multi-source domain adaptation. *British Machine Vision Conference (BMVC)* (2021)
6. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: *International Conference on Machine Learning (ICML)*. pp. 1180–1189 (2015)
7. 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 (CVPR)*. pp. 770–778 (2016)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 7132–7141 (2018)
9. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex generative adversarial network for unsupervised domain adaptation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1498–1507 (2018)
10. Hu, L., Kan, M., Shan, S., Chen, X.: Unsupervised domain adaptation with hierarchical gradient synchronization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4043–4052 (2020)
11. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. pp. 5542–5550 (2017)
12. Li, Y., Carlson, D.E., et al.: Extracting relationships by multi-domain matching. *Annual Conference on Neural Information Processing Systems (NeurIPS)* **31** (2018)
13. Li, Y., Yuan, L., Chen, Y., Wang, P., Vasconcelos, N.: Dynamic transfer for multi-source domain adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10998–11007 (2021)
14. Li, Z., Zhao, Z., Guo, Y., Shen, H., Ye, J.: Mutual learning network for multi-source domain adaptation. [arXiv:2003.12944](https://arxiv.org/abs/2003.12944) (2020), <https://arxiv.org/pdf/2003.12944.pdf>
15. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: *International Conference on Machine Learning (ICML)*. pp. 97–105 (2015)
16. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: *Annual Conference on Neural Information Processing Systems (NeurIPS)*. pp. 1640–1650 (2018)
17. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: *International Conference on Machine Learning (ICML)*. pp. 2208–2217 (2017)
18. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)* **9**(11) (2008)
19. Mallya, A., Davis, D., Lazebnik, S.: Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: *European Conference on Computer Vision (ECCV)*. pp. 67–82 (2018)
20. Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021)
21. Park, G.Y., Lee, S.W.: Information-theoretic regularization for multi-source domain adaptation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9214–9223 (2021)
22. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 1406–1415 (2019)

23. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. [arXiv:1710.06924](https://arxiv.org/pdf/1710.06924.pdf) (2017), <https://arxiv.org/pdf/1710.06924.pdf>
24. Ren, C.X., Liu, Y.H., Zhang, X.W., Huang, K.K.: Multi-source unsupervised domain adaptation via pseudo target domain. *IEEE Transactions on Image Processing (TIP)* **31**, 2122–2135 (2022)
25. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: *European Conference on Computer Vision (ECCV)*. pp. 213–226 (2010)
26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)* **15**(1), 1929–1958 (2014)
27. Wang, H., Xu, M., Ni, B., Zhang, W.: Learning to combine: Knowledge aggregation for multi-source domain adaptation. In: *European Conference on Computer Vision (ECCV)*. pp. 727–744 (2020)
28. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Tipultimodal fusion by channel exchanging. *Annual Conference on Neural Information Processing Systems (NeurIPS)* **33**, 4835–4845 (2020)
29. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3964–3973 (2018)
30. Xu, T., Ning, W., Lyu, C., Wang, K.: Joint attention-driven domain fusion and noise-tolerant learning for multi-source domain adaptation. [arXiv preprint arXiv:2208.02947](https://arxiv.org/abs/2208.02947) (2022)
31. Xu, Y., Kan, M., Shan, S., Chen, X.: Mutual learning of joint and separate domain alignments for multi-source domain adaptation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 1890–1899 (2022)
32. Zhao, H., Zhang, S., Wu, G., Moura, J.M.F., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: *Annual Conference on Neural Information Processing Systems (NeurIPS)*. pp. 8559–8570 (2018)
33. Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., Keutzer, K.: Multi-source distilling domain adaptation. In: *AAAI Conference on Artificial Intelligence (AAAI)*. pp. 12975–12983 (2020)
34. Zhu, Y., Zhuang, F., Wang, D.: Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. *Conference on Artificial Intelligence (AAAI)* pp. 5989–5996 (2019)
35. Zuo, Y., Yao, H., Xu, C.: Attention-based multi-source domain adaptation. *IEEE Transactions on Image Processing (TIP)* **30**, 3793–3803 (2021)



Edge-Guided and Cross-Scale Feature Fusion Network for Efficient Multi-contrast MRI Super-Resolution

Zhiyuan Yang¹ , Bo Zhang² , Zhiqiang Zeng³, and Si Yong Ye⁴ 

¹ School of Electronic and Information Engineering, Beihang University, Beijing, China
zyyang0416@buaa.edu.cn

² College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

³ Beijing Institute of Remote Sensing Equipment, Beijing, China

⁴ Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

Abstract. In recent years, MRI super-resolution techniques have achieved great success, especially multi-contrast methods that extract texture information from reference images to guide the super-resolution reconstruction. However, current methods primarily focus on texture similarities at the same scale, neglecting cross-scale similarities that provide comprehensive information. Moreover, the misalignment between features of different scales impedes effective aggregation of information flow. To address the limitations, we propose a novel edge-guided and cross-scale feature fusion network, namely ECFNet. Specifically, we develop a pipeline consisting of the deformable convolution and the cross-attention transformer to align features of different scales. The cross-scale fusion strategy fully integrates the texture information from different scales, significantly enhancing the super-resolution. In addition, a novel structure information collaboration module is developed to guide the super-resolution reconstruction with implicit structure priors. The structure information enables the network to focus on high-frequency components of the image, resulting in sharper details. Extensive experiments on the IXI and BraTS2020 datasets demonstrate that our method achieves state-of-the-art performance compared to other multi-contrast MRI super-resolution methods, and our method is robust in terms of different super-resolution scales. Our code is available at <https://github.com/zhiyuan-yang/Edge-Guided-Cross-Scale-MRI-Super-resolution>.

Keywords: Deep Learning · Multi-contrast Super-resolution · Cross-scale

1 Introduction

Magnetic resonance imaging (MRI) is a non-invasive and radiation-free imaging technique that plays a unique and essential role in clinical diagnosis. Compared to other imaging techniques, it can visualize anatomical tissues of different parts of the human

Z. Yang and B. Zhang—Equal Contribution

body. Despite its advantages, the acquisition of high-resolution (HR) MRI images faces challenges such as limited scanning time and patient motion [1, 2]. Therefore, MRI super-resolution (SR) has always been an important research topic in the clinical imaging community.

Traditional SR techniques such as interpolation and dictionary learning often result in over-smoothed or blurred images [3]. In recent years, deep learning (DL) based methods [4, 5] have attracted much attention, demonstrating remarkable performance. DL-based methods can be categorized into two types: single-contrast methods and multi-contrast methods. Compared to single-contrast methods, multi-contrast SR methods, which leverage complementary information from different contrasts, have shown to be more powerful. MRI routinely generates multi-contrast images with different acquisition time: T1-weighted (T1W) images normally require shorter scanning time than T2-weighted (T2W) images, so clinicians usually acquire HR T1W images (fully-sampled) and low-resolution (LR) T2W images (under-sampled). They provide complementary information about the anatomical structure, and therefore it is natural to use T1W images as the reference to acquire SR T2W images.

Reference-based SR techniques have been extensively used for both natural images and medical images [6, 7]. TTSR [8] proposes to use the hard attention mechanism to search for the most spatially relevant patch in reference images and integrate it with LR features to generate HR details. Subsequently, MASA [9] and McMRSR [10] develop a coarse-to-fine patch matching scheme that significantly reduces the computation cost and achieves better performance. In addition, WavTrans [11] incorporates the wavelet transformation into the SR framework to capture both high-frequency local structures and global information. However, these patch-based matching methods only utilize the most relevant patch in the reference images, which may overlook the complex relationship between the reference images and the LR images. Moreover, most methods adopt a straightforward approach to transfer the texture information, using either simple feature concatenation [9] or multiplication [8]. To overcome this limitation, the attention mechanism and the fully-powered transformer architecture have been introduced to extract the correlation between the LR features and the reference images. MINet [12] uses a channel-spatial attention module to fuse the features of different stages, while DCAMSR [13] proposes a dual cross-attention transformer to capture the complementary information between multi-contrast images.

Although these recent multi-contrast methods [8–13] have achieved desirable results, there are still some challenges: First, reference-based methods exclusively consider texture transfer from the reference modality, while neglecting the intrinsic anatomical structure. This neglect may lead to superficial and inconsistent SR results. In medical image analysis, it is essential to preserve the anatomical structure in the images for accurate diagnosis. Traditional methods have established the significance of incorporating structure information as a valuable prior constraint [14–17], which allows more attention to be allocated to the image details. Second, current methods only utilize texture similarities of the same scale. Earlier studies [18] have suggested that texture similarities in MRI are not only at the same scale but also across scales. Leveraging features at varying scales to aggregate the information flow can enhance the SR performance. However, simply

fusing multi-scale features may introduce more redundant noise due to the misalignment between features of different scales.

To overcome these challenges, we propose ECFNet which adopts several customized modules for SR of biomedical imaging data, depicted in Fig. 1. In particular, we adopt a coarse-to-fine feature fusion strategy to generate texture information. In addition, we add a structure branch containing high-frequency components to assist the model in generating sharper details. The proposed method effectively learns the previously neglected features of different granularities through a multi-scale feature fusion strategy and incorporates the structure information, leading to accurate SR of details. The contributions of this paper are summarized as follows: 1) We introduce the cross-scale feature fusion module (CFFM) that effectively aligns and fuses features of different scales, enhancing the aggregation of information flow. 2) The texture transfer module (TTM) is proposed to adaptively remap the distribution of reference texture with LR features so that the network can better utilize the reference information. 3) We introduce the structure information collaboration module (SICM), which facilitates interaction between features and structure information. The SICM enables the network to allocate more attention to the details while preserving the anatomical structure. Extensive experiments on two public datasets, the IXI [19] and BraTS2020 [20] datasets, demonstrate that our method achieves state-of-the-art performance.

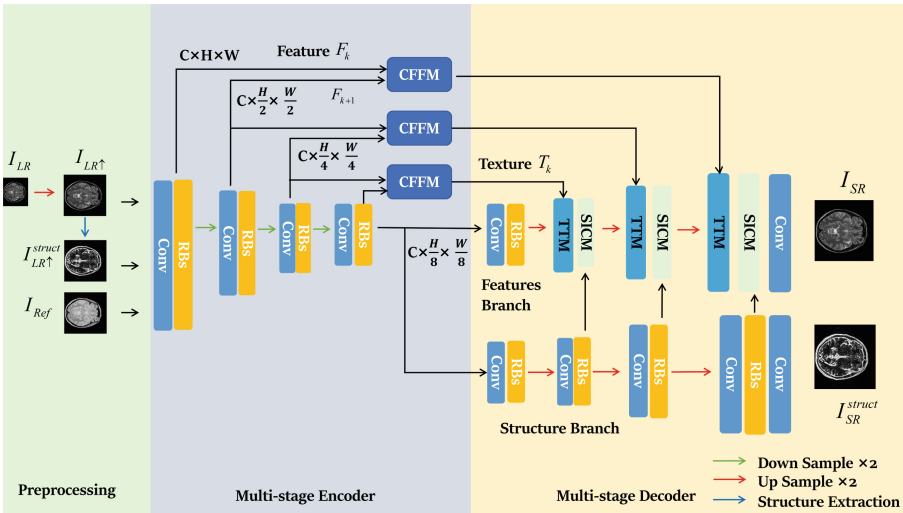


Fig. 1. The overall architecture of our proposed ECFNet.

2 Methodology

2.1 Overall Architecture

Given the LR image I_{LR} (i.e. the T2W images) and the corresponding reference image (Ref) I_{Ref} (i.e. the T1W images), the ECFNet can accurately restore I_{LR} to the SR image I_{SR} . As shown in Fig. 1, the framework mainly consists of three parts: the preprocessing, the multi-stage encoder, and the multi-stage decoder. In the preprocessing stage, I_{LR} is first interpolated to the same size as I_{Ref} , and the Sobel operator is used to extract the edge map $I_{LR\uparrow}^{struct}$. The multi-stage encoder contains four layers, where each layer consists of a down-sample convolutional layer and residual blocks. After passing $I_{LR\uparrow}$ and I_{Ref} into the encoder, features with different scales are obtained, denoted as F_k and F_k^{Ref} where $k = 1, 2, 3, 4$. The CFFM aligns and fuses the features extracted from I_{Ref} and $I_{LR\uparrow}$ to generate the coarse-to-fine texture T_k . In the multi-stage decoder, the texture is first aggregated with the features using the TTM. After that, the SICM facilitates interaction between the features and structure information, refining the details according to the structure information. Finally, we obtain the SR image I_{SR} and the SR structure map I_{SR}^{struct} using a simple convolutional layer.

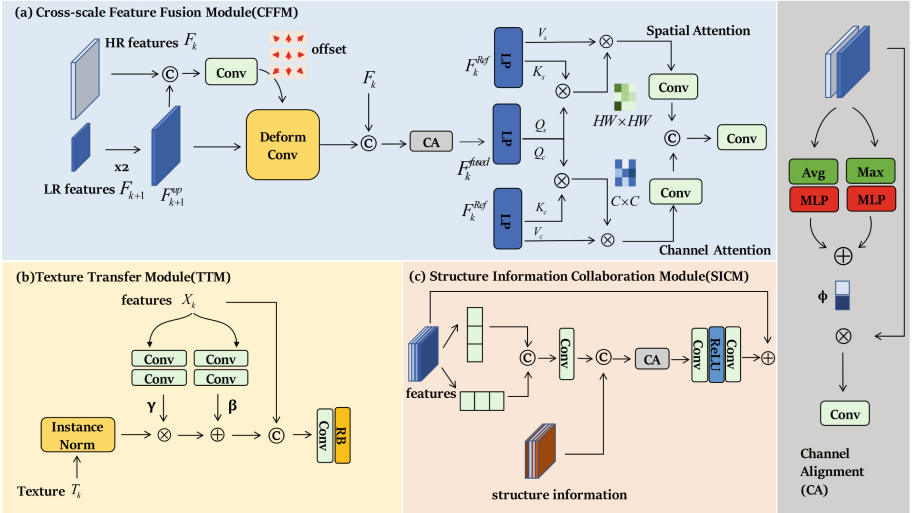


Fig. 2. The components of ECFNet: (a) Cross-scale Feature Fusion Module (CFFM); (b) Texture Transfer Module (TTM); (c) Structure Information Collaboration Module (SICM).

2.2 Cross-Scale Feature Fusion Module

In CFFM shown in Fig. 2 (a), LR features F_k are aligned and then fused with Ref features F_k^{Ref} to incorporate the information from reference images. Since cross-scale similarities of features are widespread, utilizing the aggregated information from different scales can

improve the SR results. The key issue is that misalignment of position and channel may appear across different layers, which hinders the comprehensive integration of multi-scale information. Therefore, we propose to adaptively align the up-sampled LR features F_{k+1}^{up} with HR features F_k .

First, deformable convolution [21] is used to introduce learnable offsets to the spatial sampling locations, augmenting the alignment of the receptive field in the down-sampled features with HR features. The offset is learned by convolutional layers from concatenated features:

$$offset = Conv_{3 \times 3}(Concat(F_k, F_{k+1}^{up})). \quad (1)$$

The deformable convolution then uses the offset to get the aligned features:

$$F_{k+1}^{aligned}(p_i) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot F_{k+1}^{up}(p_i + p_n + \Delta p_n). \quad (2)$$

where p_i denotes a pixel in aligned features $F_{k+1}^{aligned}$, $w(\cdot)$ and Δp_n are the weight and the offset respectively. The aligned LR features $F_{k+1}^{aligned}$ are subsequently concatenated with the HR features F_k . To alleviate the channel misalignment among features of different scales, we introduce a channel alignment (CA) module. The global max pooling layer and global average pooling layer are used to extract the channel information from the concatenated features F_k^{concat} respectively, and the outcome is denoted as $P_{avg} \in \mathbf{R}^{C \times 1 \times 1}$ and $P_{max} \in \mathbf{R}^{C \times 1 \times 1}$. A multi-layer perceptron (MLP) consisting of two fully connected layers with a reduction rate of 16 is then used to get the channel alignment coefficient ϕ . The output is obtained by:

$$F_k^{fused} = \phi \cdot F_k^{concat} + F_k^{concat}. \quad (3)$$

After aggregating the features at different scales, a dual cross-attention transformer [13] is used to generate reference texture by utilizing the complementary information from F_k^{Ref} . Linear projection functions are used to compute the query, value, and key of the features, and the spatial and channel attention are then obtained by

$$T_s = softmax\left(\frac{Q_s \times K_s^T}{\sqrt{d}} \times V_s\right), \quad (4)$$

$$T_c = softmax\left(\frac{Q_c \times K_c^T}{\sqrt{d}} \times V_c\right). \quad (5)$$

Finally, the spatial and channel attention are concatenated and reduced to half channel with depth-wise convolution. The obtained features are then processed by the residual blocks to generate textures T_k .

2.3 Multi-stage Decoder

At each stage of the decoder, the extracted texture is first integrated with the features using TTM, where the distribution of the texture is remapped with the features. The

details are then refined according to the structure information using the SICM so that more attention is allocated to them.

Since the distribution of the extracted texture may be inconsistent with LR features, simple concatenation may lead to suboptimal results. Inspired by the work of [9], we design a texture transfer module (TTM) as shown in Fig. 2 (b) to remap the distribution of the texture with LR features. The instance normalization is used to extract the structure of texture and discard its style:

$$T_k \leftarrow \frac{T_k - \mu_{T_k}}{\sigma_{T_k}}. \quad (6)$$

After that, the affine transformation is used to update the features:

$$T_k \leftarrow X_k \otimes \beta + \gamma. \quad (7)$$

Two separate convolutional blocks are used to learn the affine transformation parameters β and γ so that the features can adapt the style to the texture while maintaining the structure. Then the transferred texture is concatenated with the features and fused by a residual block. Compared to simple multiplication or concatenation, the TTM takes characteristics of both features and texture into consideration. The adaptive fusion process can enhance the incorporation of reference information.

MRI has a large plain background and small important target areas. These areas contain rich tissue information that is important for accurate diagnosis. The edge map corresponds to the high-frequency components in the images, therefore incorporating the edge information can guide the network to allocate more attention to the details during the SR reconstruction. Since the edge map has zero values in most areas, an asymmetric convolutional group consisting of 1×3 and 3×1 convolutions is used to extract geometric structure both vertically and horizontally, and 1×1 convolution is adopted to refine the features:

$$X_k^{edge} = Conv_{1 \times 1}(Concat[Conv_{3 \times 1}(X_k), Conv_{1 \times 3}(X_k)]). \quad (8)$$

The channel alignment (CA) module is adopted to remap the distribution of structure information with the features. Next, to improve the stability of the network training, we use the residual connection to get the fused features:

$$X_{k-1} = Conv(\text{ReLU}(Conv(X_k^{aligned}))) + X_k. \quad (9)$$

The SICM makes the network easier to preserve the anatomical information for accurate SR and leads to sharper details.

2.4 Loss Function

The L_1 loss is used for the reconstruction and structure loss. The total loss function is

$$L = \frac{1}{N} \sum_{n=1}^N L_1(I_{SR}, I_{HR}) + L_1(S(I_{HR}), I_{SR}^{struct}), \quad (10)$$

where $S(\cdot)$ represents the Sobel operator.

3 Experiments

Datasets and Baselines. The IXI [19] and BraTS2020 [20] datasets are used to evaluate our proposed method. The IXI dataset contains registered T2W and proton density weighted (PDW) 3D MRI volumes of 578 subjects, and we use the PDW MRI as the reference modality. We adopt the same preprocessing procedure of [22], where 3D volumes are clipped into the size of $240 \times 240 \times 96$. 500 subjects are randomly selected as the training set and another 70 subjects as the testing set. For each subject, 10 slices are selected. 2-fold and 4-fold down-sampled T2W LR images are created using the k-space truncation. The BraTS2020 dataset contains 369 subjects for the training dataset and 125 subjects for the validation dataset. Each subject has 4 modalities with size of $240 \times 240 \times 155$, and we use T1W images as reference images. 300 subjects are randomly chosen for training and another 100 subjects for testing. We compare our methods with four multi-contrast methods (MINet [12], DCAMSR [13], TTSR [8], WavTrans [11]) and a single-contrast method (SwinIR [23]). Peak signal-to-noise ratio (PSNR) and structure similarity index measure (SSIM) are used to evaluate the performance of different methods.

Table 1. Quantitative results on two datasets with different scales. Red numbers indicate the best result, and blue numbers indicate the second-best result.

Dataset	IXI				BraTS2020			
	2×		4×		2×		4×	
Metric	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TTSR	38.487	0.981	30.400	0.920	39.597	0.990	32.671	0.961
SwinIR	37.002	0.977	29.250	0.908	39.418	0.991	31.758	0.956
MINet	39.925	0.984	34.093	0.942	40.315	0.992	33.602	0.965
DCAMSR	40.324	0.986	35.908	0.967	40.175	0.991	33.806	0.967
WavTrans	39.719	0.981	33.443	0.963	39.857	0.995	33.406	0.971
Ours	41.823	0.987	37.213	0.970	41.218	0.995	34.985	0.972

Implementation Details. We train all the models on NVIDIA GeForce RTX 3090 GPUs. Our model is trained using the Adam optimizer with a learning rate of $2e-4$ for 50 epochs. The batch size is set as 10. The parameters of the Adam optimizer, α and β , are set to 0.9 and 0.999 respectively. All the compared models are trained using their default parameter settings.

Quantitative Results. Table 1 summarizes the PSNR and SSIM scores on two public datasets in 2-fold and 4-fold SR. Compared with other methods, our method achieves the best results in all cases, which proves the effectiveness of our method. In the challenging 4-fold SR situation, our method can still achieve a desirable result. We give the multi-feature fusion strategy credit for it. It aggregates information flow from different scales so that even in the LR situation, the network is still able to extract effective texture information. Besides, the alignment procedure can effectively reduce the redundant noise when fusing different scale features.

Table 2. Ablation study on the IXI dataset with 4-fold SR.

Variant	Modules			Metrics	
	CFFM	TTM	SICM	PSNR	SSIM
<i>w/o</i> multi-scale feature alignment	×	✓	✓	33.478	0.941
<i>w/o</i> texture transfer	✓	×	✓	35.437	0.968
<i>w/o</i> structure branch	✓	✓	×	35.312	0.965
full version	✓	✓	✓	37.213	0.970

Qualitative Results. Figure 3 shows the SR results and the corresponding error maps on two datasets with different SR rates. In the error maps, prominent features indicate poor detail reconstruction. It can be observed that our method is superior compared to other methods in both datasets, which proves the robustness of our method. Furthermore, our method generates sharper texture details compared to other methods because we incorporate the structure information, allowing the network to focus on the details during the reconstruction process. In the SR process, the features are adaptively adjusted in the informative regions guided by the structure information, resulting in more details.

Ablation Study. We conduct ablation studies on the IXI dataset for 4-fold SR to evaluate the effectiveness of different modules within our framework, and the results are shown in Table 2. Three variant networks are used: 1) *w/o* multi-scale feature alignment, where the cross-scale alignment part in the CFFM is not used. 2) *w/o* texture transfer, which is our model without the TTM. 3) *w/o* structure branch, which is our model without the edge map extraction and the edge branch. The results indicate that the variant *w/o* multi-scale feature alignment performs worst, which proves that our alignment module effectively integrates features from different scales. The degradation of variant *w/o* struct branch is consistent with our conclusion that structure information can enhance the SR reconstruction and lead to sharper details. Furthermore, the improvement from the variant *w/o* TTM to the full version also proves the effectiveness of the texture transfer module.

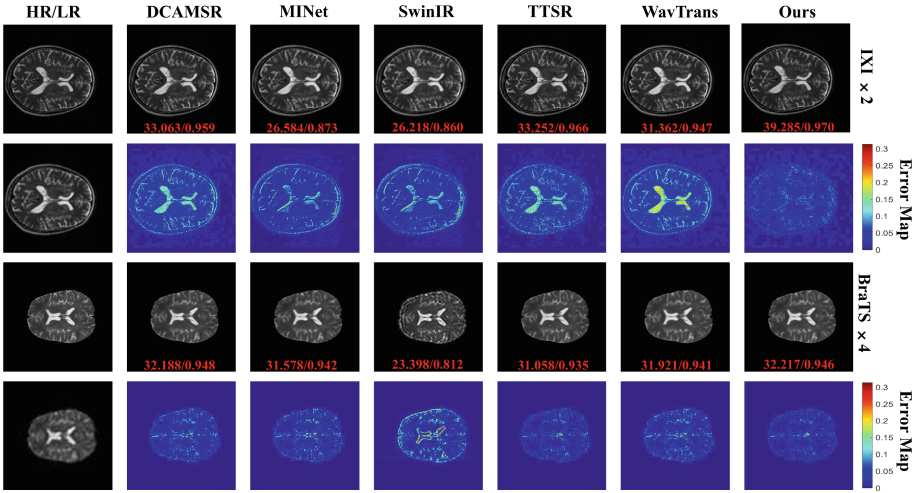


Fig. 3. Qualitative results and error maps of different methods on two datasets. The first/third row are the SR results, and the second/fourth row are the corresponding error maps. The brighter color suggests more errors.

4 Conclusion

In this study, we propose an edge-guided and cross-scale feature fusion network for multi-contrast MRI super-resolution. Specifically, we design a novel pipeline to utilize cross-scale similarities in MRI that can provide comprehensive information. In addition, we incorporate the structure information to guide the network towards generating sharper textures. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance, especially in the challenging four-fold SR. Our work provides a possible direction for further research in processing multi-contrast MRI, which has great potential uses in many medical applications. In the future, we would like to explore multi-contrast MRI super-resolution at arbitrary scales.

Acknowledgments. This project is supported by the Lee Kong Chian School of Medicine - Ministry of Education Start-Up Grant.

Disclosure of Interests. Authors have no conflict of interest to declare.

References

1. Gordillo, N., Montseny, E., Sobrevilla, P.: State of the art survey on MRI brain tumor segmentation. *Magn. Reson. Imaging.* 31, 1426–1438 (2013)
2. Despotović, I., Goossens, B., Philips, W.: MRI Segmentation of the Human Brain: Challenges, Methods, and Applications. *Comput. Math. Methods Med.* 2015, 1–23 (2015)
3. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image Super-Resolution Via Sparse Representation. *IEEE Trans. Image Process.* 19, 2861–2873 (2010)

4. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In: Leal-Taixé, L. and Roth, S. (eds.) ECCV 2018 Workshops, LNCS, vol. 11133, pp. 63–79. Springer, Cham (2019)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 295–307 (2016)
6. Rousseau, F.: Brain Hallucination. In: Forsyth, D., Torr, P., and Zisserman, A. (eds.) ECCV 2008, LNCS, vol. 5302, pp. 497–508. Springer, Heidelberg (2008)
7. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, pp. 327–340. ACM, Los Angeles (2001)
8. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5790–5799, IEEE (2020)
9. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: MASA-SR: Matching Acceleration and Spatial Adaptation for Reference-Based Image Super-Resolution, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6364–6373. IEEE (2021)
10. Li, G., Lv, J., Tian, Y., Dou, Q., Wang, C., Xu, C., Qin, J.: Transformer-empowered Multi-scale Contextual Matching and Aggregation for Multi-contrast MRI Super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20604–20613. IEEE, New Orleans (2022)
11. Li, G., Lyu, J., Wang, C., Dou, Q., Qin, J. WavTrans: Synergizing Wavelet and Cross-Attention Transformer for Multi-contrast MRI Super-Resolution. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, LNCS, vol. 13436, pp. 463–473. Springer, Cham (2022)
12. Feng, C., Fu, H., Yuan, S., Xu, Y. Multi-contrast MRI Super-Resolution via a Multi-stage Integration Network. In: de Bruijne, M., et al. (eds.) MICCAI 2021, LNCS, vol. 12906, pp. 140–149. Springer, Cham (2021)
13. Huang, S., Li, J., Mei, L., Zhang, T., Chen, Z., Dong, Y., Dong, L., Liu, S., Lyu, M.: Accu-rate Multi-contrast MRI Super-Resolution via a Dual Cross-Attention Transformer Network. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., and Taylor, R. (eds.) MICCAI 2023, LNCS, vol. 14229, pp. 313–322. Springer Cham (2023)
14. Zheng, H., Zeng, K., Guo, D., Ying, J., Yang, Y., Peng, X., Huang, F., Chen, Z., Qu, X.: Multi-Contrast Brain MRI Image Super-Resolution With Gradient-Guided Edge Enhancement. *IEEE Access.* 6, 57856–57867 (2018)
15. Jafari-Khouzani, K.: MRI upsampling using feature-based nonlocal means approach. *IEEE Trans. Med. Imaging.* 33, 1969–1985 (2014)
16. Rousseau, F.: A non-local approach for image super-resolution using intermodality priors. *Med. Image Anal.* 14, 594–605 (2010)
17. Han, S., Remedios, S., Carass, A., Schär, M., Prince, J.L.: MR Slice Profile Estimation by Learning to Match Internal Patch Distributions. *IPMI 2021*, LNCS, vol. 12729, pp. 108–119. Springer Cham (2021)
18. Plenge, E., Poot, D.H.J., Niessen, W.J., Meijering, E.: Super-Resolution Reconstruction Using Cross-Scale Self-similarity in Multi-slice MRI. In: Mori, K., Sakuma, I., Sato, Y., Barrillot, C., and Navab, N. (eds.) MICCAI 2013, LNCS, Part III. pp. 123–130. Springer, Heidelberg (2013)
19. IXI Dataset, <https://brain-development.org/ixi-dataset/>, last accessed 2023/12/19
20. Multimodal Brain Tumor Segmentation Challenge 2020: Data, <https://www.med.upenn.edu/cbica/brats2020/data.html>, last accessed 2023/12/19
21. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable Convolutional Networks. In: 16th IEEE International Conference on Computer Vision, ICCV 2017, pp. 764–773. IEEE, Venice (2017)

22. Zhao, X., Zhang, Y., Zhang, T., Zou, X.: Channel Splitting Network for Single MR Image Super-Resolution. *IEEE Trans. Image Process.* 28, 5649–5662 (2019)
23. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image Restoration Using Swin Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1833–1844. IEEE (2021)



Dual-ResShift: Dual-Input Separated Features Residual Shift Diffusion Model for CTA Image Super-Resolution

Feng Jiang, Jing Wen^(✉), and Yi Wang

College of Computer Science, Chongqing University, Chongqing 400030, China
wj@cqu.edu.cn

Abstract. Computed tomography angiography (CTA) scans provide doctors with an accurate visualization of the vascular system that helps them make diagnostic decisions. To help doctors make a quick diagnosis, the resolution of CTA images needs to be improved. With the advancement of deep learning technology, many neural network models have been proposed to improve image quality. Most current image super-resolution (SR) methods still have problems, such as texture loss and boundary-blurring, and there are few methods for CTA images. In this paper, we propose the Separation Feature Residual Diffusion Model (Dual-ResShift) for generating high-resolution (HR) CTA images with only 15 sampling steps. In the model, we propose a feature-separated SFUNet and a new feature extraction algorithm, GBVS-Enhanced based on Graph-based visual saliency (GBVS), to enhance the extraction of high-frequency features. In addition, we design a loss function named \mathcal{L}_{Edge} to guide the diffusion model to optimize the image. Our method was verified on the clinical CTA dataset and AVT dataset. The PSNR reached 26.4165, SSIM reached 0.7440, and LPIPS reached 0.0484 on the clinical CTA dataset. The PSNR reached 28.6791, SSIM reached 0.7805, and LPIPS reached 0.0365 on the AVT dataset. Experiments show that Dual-ResShift can outperform existing methods. Our code and model are put on <https://github.com/prefectmoon/Dual-ResShift-code>.

Keywords: Super resolution · Computed tomography angiography · Residual diffusion

1 Introduction

Computed tomography angiography (CTA) is a crucial radiological technique in visualizing and diagnosing cerebrovascular diseases [17]. Despite the benefits of noninvasiveness and 3D imaging presentation, CTA's resolution remains relatively low, and the complexity of the edges of blood vessel pixels increases during

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78398-2_15.

super-resolution (SR) processing. The challenge of attaining sub-millimeter-level ultra-high-resolution (HR) images in CTA scans is increasingly prominent [31]. Therefore, improving the clarity of CTA images becomes essential to enhance diagnostic accuracy. Image SR involves restoring HR images from low-resolution (LR) counterparts, serving as a critical image processing technique in computer vision [35]. Leveraging this method effectively improves image quality, assisting physicians in enhancing diagnostic accuracy when utilizing CTA images.

Traditional SR methods often suffer from the loss of texture information, yielding SR images that lack realism. In contrast, deep learning-based SR methods excel in extracting the nonlinear relationship between HR and LR, resulting in superior high-frequency information preservation. Many SR methods based on deep learning have achieved remarkable results [5, 7, 8]. Early approaches typically acquired LR images under known degradation processes, lacking generalization and potentially yielding unsatisfactory results in practical applications; there can be many visual artifacts. Based on unknown degradation, works such as BSRGAN [42], Real-ESRGAN [34], and SwinIR [18] generate SR, better simulating real-world scenarios. Recently, diffusion models have been introduced to SR tasks, showcasing outstanding generative capabilities and attracting extensive attention. Applying the diffusion model to image SR has also yielded promising results, indicating significant potential in SR tasks.

Although the existing method of SR based on the diffusion model has made a series of achievements, it has three main shortcomings: Firstly, Many methods ignore the texture information of LR images and do not retain valid structure information in the reconstruction process. Many unpleasant visual artifacts appear in the generated images, such as joint misalignment or texture distortion. Secondly, these methods mainly rely on real-world natural image datasets, such as set5 [2], DIV8K [11], Flickr2K [1], etc., and may not achieve satisfactory results when applied to CTA images. The generation of SR images based on the diffusion model generally requires many sampling steps (50 to more than 1000 steps) to obtain high-quality images after a long inference, it is computationally expensive, and not suitable for real-time applications. In addition, the diffusion model has not been applied to blind SR of CTA images. Therefore, we construct Dual-ResShift based on the ResShift [41] designed for SR to address these challenges. Our main contributions are as follows:

1. We design a dual-input SFUNet for separating features to help the model obtain a more comprehensive feature mapping.
2. Based on the characteristics of CTA images and the Graph-based visual saliency (GBVS) algorithm, we present a GBVS-Enhanced algorithm to process high-frequency information in LR images, which provides more texture information for the model.
3. We propose a new loss function named \mathcal{L}_{Edge} , which helps optimize the model and makes the generated image more consistent with the target image on edge information.

Our experimental results show that Dual-ResShift can generate high-quality clear images from blindly degraded LR, exceeding the performance of most current models.

2 Related Work

2.1 Image Super-resolution

The earliest model to achieve SR using deep neural networks is SRCNN[7], which utilizes a three-layer convolutional network to fit nonlinear mappings and obtain HR images. Subsequent methods such as ESPCN [27], EDSR [19], RCAN [44], SRGAN [15], etc., effectively improve the resolution of the image. [3] to improve the resolution of medical images by using multiple improved residual networks is presented. [9] proposed a 3D deep dense connection neural network to improve the quality of brain MRI scanning. [21] presented a problem of using deep residual networks to improve the lack of correlation between feature information in CT images.

Early works involved training LR images obtained through bicubic downsampling or K-space truncation. These methods lack generalization, which greatly underestimates the complexity of real noise, and the effect will be reduced in practical application scenarios. Some studies utilize blind SR methods to provide effective degradation pipelines, where the image is randomly degraded from the Angle of noise and fuzzy kernel, such as BSRGAN [42] and Real-ESRGAN [34]. [45] realized blind MRI oversampling based on CycleGAN [46] architecture, which reduced image distortion. [30] proposed a parallel alternate iterative optimization degradation strategy and enhanced spatial feature transform residuals to implement the blind SR method for CMRI images. [26] have improved the regression process and ESRGAN’s loss function to reconstruct MRI images and have good generalization.

2.2 Diffusion Model for Image Super-resolution

DDPM [13] exceeds GAN [10] in many generation tasks by continuously adding Gaussian noise destruction training data in a forward process and then reversing noise recovery data in a reverse process. Subsequently, proposed models such as score-based generative model [29], DDIM [28], guide diffusion [6], LDM [23]etc., have been further developed based on DDPM, reducing sampling steps and improving the generation quality of diffusion models. Due to the excellent generative performance of diffusion models, some studies have also been exploring SR reconstruction based on diffusion models. For example, SRDiff [16], SR3 [25], DiffIR [40], StabeSR [32], and ResShift [41]. However, the application of diffusion models in medical imaging is limited. [39] by integrating the self-attention mechanism into DDPM, a new deep learning SR framework for brain MRI images based on DDPM is proposed. InverseSR [33] has achieved SR for MRI images based on their sparsity. At the same time, DisC-Diff [20] combined T1 and T2 modalities of MRI images for multi-contrast SR reconstruction.

The diffusion model can produce a stable optimization process and has shown excellent performance in SR reconstruction work, but blind SR CTA image reconstruction based on the diffusion model has not been realized. Due to the complex texture of CTA images, edge blurring may lead to unsatisfactory results, and feature complexity in CTA slices poses challenges for diffusion model learning of images.

3 Methodology

3.1 Overall Architecture

The diffusion component proposed in Dual-ResShift builds upon ResShift [41], establishing a Markov chain by manipulating residuals between HR and LR images. Compared to other contemporary diffusion models, it requires only 15 sampling steps and boasts highly flexible noise control mechanisms. This allows for precise adjustment of residual and noise levels during transitions. As depicted in Fig. 1, here is a Markov chain between HR and LR obtained by shifting residuals. HR x_0 adds Gaussian noise in 15 steps to get x_{15} in the forward diffusion process. The reverse process denoises x_{15} with LR and GBVS-Enhanced as conditions to get the SR image.

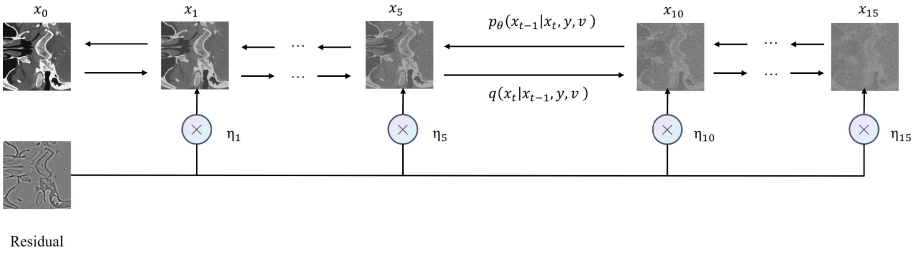


Fig. 1. Residual diffusion of Dual-ResShift.

Forward Process Given HR and LR image pairs, let the residual between HR and LR be represented as e_0 , where $e_0 = y_0 - x_0$. The direction of forward process optimization is to gradually adjust the residual e_0 through a Markov chain of length T to transform x_0 into y_0 . In the implementation process, an offset sequence $\eta_{t=1}^T$ is introduced, which monotonically increases with time t and satisfies $\eta_1 \rightarrow 0, \eta_T \rightarrow 1$. We denote κ as a hyper-parameter controlling the noise variance, the noise level in x_t is proportional to η_t , and κ is the scale factor. I is the identity matrix. In addition, the marginal distribution at any timestep t is analytical, so the forward process is distributed as follows Eq. 1:

$$q(x_t|x_0, y_0) = N(x_t; x_0 + \eta_t + e_0, \kappa^2 \eta_t I), t = 1, 2, \dots, T \tag{1}$$

Reverse Process The reverse process aims to generate a new SR image from x_T . This is accomplished by constructing the reverse distribution $p_\theta(x_{t-1}|x_t, y_0, v_0)$, conditioned on its associates LR image and the feature map of CTA image obtained by GBVS-Enhanced algorithm. This reverse process can be expressed as follows Eq. 2:

$$p(x_0|y_0, v_0) = \int p(x_T|y_0, v_0) \prod_{t=1}^T p_\theta(x_{t-1}|x_t, y_0, v_0) dx_{1:T}$$

$$p(x_T|y_0, v_0) \approx N(x_T|y_0, v_0, \kappa^2 I)$$

$$p_\theta(x_{t-1}|x_t, y_0, v_0) = N(x_{t-1}; \mu_\theta(x_t, y_0, v_0, t), \sum_{\theta} (x_t, y_0, v_0, t)) \quad (2)$$

where $p_\theta(x_{t-1}|x_t, y_0, v_0)$ represents the reverse kernel from x_t to x_{t-1} with a learnable parameter θ . We show the overall structure of Dual-ResShift in Fig. 2.

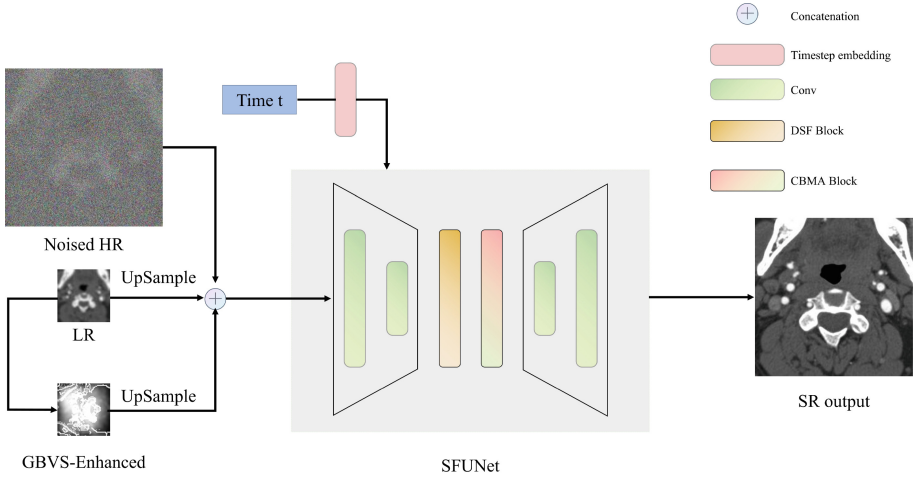


Fig. 2. The overall structure of Dual-ResShift.

3.2 Separated Features Unet

It isn't easy to obtain the inverse process distribution, so we incorporate the SFUNet for separating features and GBVS-Enhanced as a second input to supplement the texture information. As shown in Fig. 3, SFUNet derived the distribution of the inverse process by learning CTA images with separated features, combined with the proposed $\mathcal{L}_{Total} = \mathcal{L}_{MSE} + \mathcal{L}_{SF} + \mathcal{L}_{MS-SSIM} + \mathcal{L}_{Edge}$ joint loss function to optimize the model, and guided the model to generate new HR images conditionally.

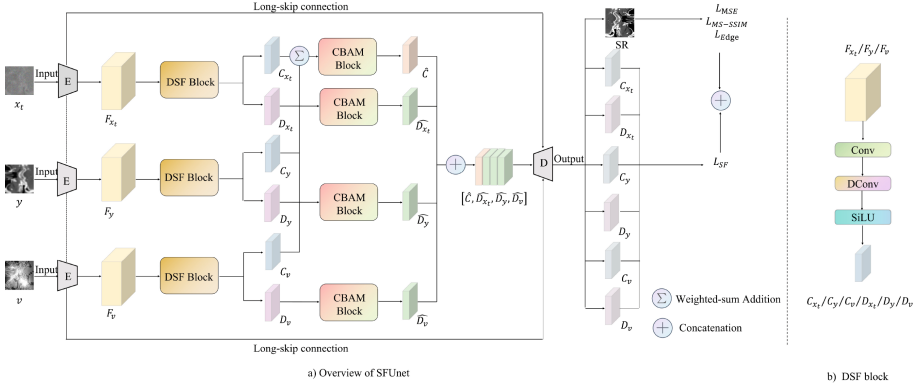


Fig. 3. The structure of SFUNet.

The SFUNet has three encoders that extract three features separately. We introduce a deformable feature separation(DSF) block, which includes a convolutional layer, a deformable convolutional layer, and a sigmoid linear unit (SiLU) activation function. The offset introduced by the deformable convolution [4] in the receptive field can better fit the irregular edges of blood vessels and bones in the image. The input HR with added noise as x_t , LR as y , and the feature map obtained by GBVS-Enhanced as v are sent into three identical encoders to extract features, yielding F_{x_t} , F_y , and F_v , respectively. The DSF block processes these three features to yield common features C_{x_t} , C_y , and C_v , alongside distinct features D_{x_t} , D_y , and D_v . The derived common features C_{x_t} , C_y , and C_v are weighted and combined using the CBAM [38] block to produce the composite feature C_{all} . Concurrently, the independent features D_{x_t} , D_y , and D_v are separately weighted and inputted into the CBAM module to derive \hat{D}_{x_t} , \hat{D}_y , and \hat{D}_v , signifying the independent features of the three inputs. The convolutional layer and SiLU activation function are used to integrate the common features and three different features and output a set of weights $[\hat{D}_{x_t}, \hat{D}_y, \hat{D}_v, \hat{C}]$. Finally, the weights are sent to the decoder for upsampling. The model outputs include the common feature of HR, the distinct feature, and the predicted SR.

3.3 GBVS-Enhanced

Through some experiments, we have found that current SR methods exhibit issues with unclear textures in CTA data. Some methods generate information that does not exist in the HR images, which can be highly dangerous for medical diagnosis. To solve this problem, we adopt a dual-input approach to provide more information to the network. The network’s inputs consist of the anxious HR image, the LR image obtained using Real-ESRGAN blind SR, and the GBVS-Enhanced extracted from the LR image. The GBVS-Enhanced algorithm is described as follows [12]: First, the R , G , B , and $L = \max(R, G, B)$ channels of the LR image are processed using a Gaussian pyramid to obtain

four pyramids: R , G , B , and L . Then, color features are extracted by computing $L = \max(R, G, B)$, $CBY = (B - \min(R, G))/L$, and $CRG = (R - G)/L$. CBY and CRG represent the differences in the blue-yellow and red-green components of the color features, respectively. Next, by applying the Gabor filter to L feature mapping to extract directional features, the feature response M of L image in different directions is obtained. Finally, for the feature map M , a weighted edge with weight ω is introduced between any two nodes (i, j) and (p, q) as (i, j) to (p, q) edge. The formula is expressed as follows Eq. 3:

$$\begin{aligned} \omega((i, j), (p, q)) &= A((i, j) || (p, q)) \cdot F(i - p, j - q) \\ A((i, j) || (p, q)) &= \left| \log \frac{M(i, j)}{M(p, q)} \right| \\ F(i - p, j - q) &= \exp\left(-\left(\frac{(i - p)^2 + (j - q)^2}{2\sigma^2}\right)\right) \end{aligned} \quad (3)$$

Where σ is the parameter, $M(i, j)$ and $M(p, q)$ are the eigenvalues of $(i, j), (p, q)$. A Markov chain with a normalized weight ω is defined on M . A is obtained by the stable state of the Markov chain on M , and then A is multiplied by the fully connected graph F to obtain the final salient graph. To enhance the contrast between blood vessels and surrounding tissues in the CTA image, we fuse the LR color values (obtained by $(R + G + B)/3$) and the LR edge feature map with the GBVS feature map as our second input. As shown in Fig. 4, the GBVS-Enhanced feature map highlights the most important information in the image, making it more prominent, and adding color value and edge feature map makes the texture information more abundant.

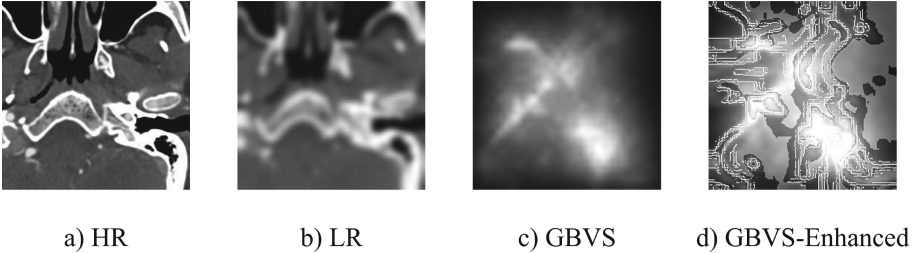


Fig. 4. Comparison of HR image, LR image, GBVS algorithm, and the proposed GBVS-Enhanced algorithm.

3.4 Loss Function

The loss function can guide the CTA image generated by the model to be closer to HR. In order to effectively learn features from the image and help the model converge, we use the joint loss function in Dual-ResShift. In addition, we denote

the target as $X = x_0$, the output of the model as $Y = f_{\theta}(x_t, y_0, v_0, t)$. The formula is expressed as Eq.4.

$$\mathcal{L}_{Total} = \mathcal{L}_{MSE} + \mathcal{L}_{SF} + \mathcal{L}_{MS-SSIM} + \mathcal{L}_{Edge} \quad (4)$$

MSE Loss The diffusion model predicts x_0 . To help the model converge, we use the \mathcal{L}_{MSE} loss function. It is expressed as Eq.5:

$$\mathcal{L}_{MSE} = \frac{1}{n} \sum_{i=1}^n |X_i - Y_i|^2 \quad (5)$$

Separated Features Loss In SFUNet, for the features of the model to minimize the differences between shared features and maximize the differences between independent features, we calculate the $L2$ distance between shared and independent features and take their ratio as the optimization Eq. 6:

$$\mathcal{L}_{SF} = \frac{\|C_{x_t} - C_y\|_2 + \|C_{x_t} - C_v\|_2 + \|C_y - C_v\|_2}{\|D_{x_t} - D_y\|_2 + \|D_{x_t} - D_v\|_2 + \|D_y - D_v\|_2} \quad (6)$$

MS-SSIM Loss To make the overall structure information of the SR CTA image more complete and ensure that the image is not only close to the original image at the pixel level but also consistent in structure, we introduce Eq. 7 calculation of similarity on multiple scales [37].

$$\mathcal{L}_{MS-SSIM} = 1 - [l_m(X, Y)]^{\alpha M} \times \prod_{j=1}^M [c_j(X, Y) \times s_j(X, Y)]^{\alpha j} \quad (7)$$

The original scale of the image is 1, and the maximum scale is M, and $j \in [1, M]$, l_m , c_j , and s_j are represented by brightness, contrast, and structural similarity.

Edge Loss The blood vessel edges in CTA images are complex. We designed an \mathcal{L}_{Edge} to improve the edge freshness in CTA images. First, we convert both the model's output and ground truth images into grayscale. Then, we calculate the difference between the maximum and minimum filters and extract the edge information. The difference between the model's output and ground truth images is then measured by calculating the $L1$ loss between these edge images. We denote the edge from X as E_X , the edge from Y as E_Y , and the complete formula is expressed in Eq. 8. We show four samples of extracted edge features in Fig. 5.

$$\mathcal{L}_{Edge} = \frac{1}{n} \sum_{i=1}^n |E_{X_i} - E_{Y_i}| \quad (8)$$

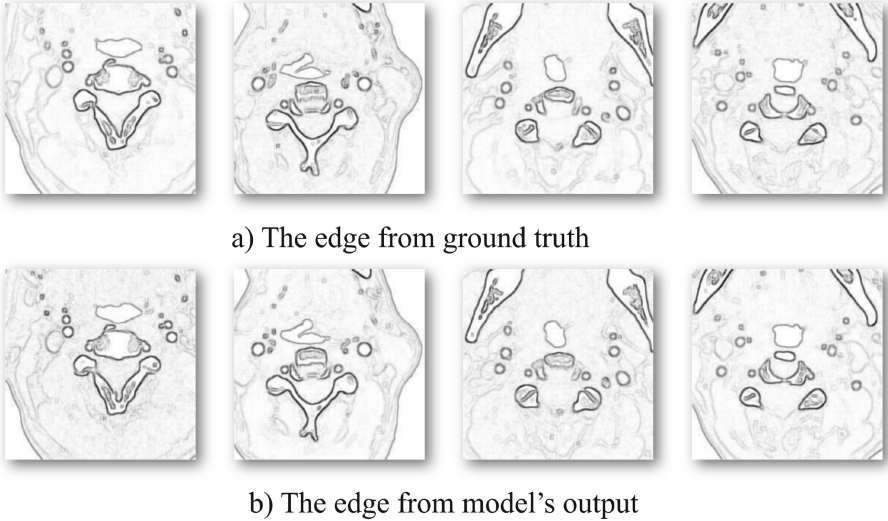


Fig. 5. The examples of the edge images.

4 Experiments

4.1 Datasets

We ran experiments on two datasets of CTA. The data set was derived from CTA images and related electronic medical records of 200 patients with intracranial aneurysms provided by Xinqiao Hospital of Chongqing Army Military Medical University and Banan Hospital Affiliated with Chongqing Medical University. We used 147 sets as a training set, 10 sets as a validation set, and 20 sets as a test set. The other data set from the publicly available Aortic Vessel Tree (AVT) [22], of which we used 22 sets as the training set, 3 sets as the validation set, and 8 sets as the test set.

4.2 Experimental Settings

The Dual-ResShift, implemented based on PyTorch, follows the UNet architecture and settings described in guide diffusion [6]. It incorporates 2 BigGAN residual blocks and utilizes 4 attention heads with attention mechanisms employed at resolutions of 8×8 , 16×16 , and 32×32 . During training, the batch size is set to 64, Adam optimizer is used with a learning rate of $5e-5$, and the training is carried out for 100k iterations on an NVIDIA RTX H800 80G GPU. Regarding the hyperparameters, T is set to 15, k to 2.0, and p to 0.3.

In the training data, we cut 30 256×256 slices of CTA images of the training set and verification set from the Z-axis direction. In the test data set, we clipped 10 256×256 slices of CTA images from the Z-axis direction (due to the small number of images in the AVT dataset, we used 30 256×256 for our test set),

and the RealESRGAN [34] was also used to obtain the LR images for the test set degradation model. However, to better simulate the LR situation in reality, we delete the second-order operation based on RealESRGAN degradation, set the probability of randomly sampled isotropic Gaussian kernel and anisotropic Gaussian kernel to [0.6,0.4], and set the kernel width of isotropic Gaussian kernel from [0.2,0.8]. Anisotropic Gaussian kernels' width along the x and y axes is randomly sampled from [0.2,0.8]. For the added Gaussian noise, we randomly select from the levels [1, 15] and Poisson noise from [0.05,0.3]; the ratio of Gaussian noise and Poisson noise are [0.5,0.5]. The downsampling of images from 256*256 to 64*64 is implemented randomly from "area," "bilinear," and "bicubic." Finally, the image quality is compressed using JPEG, and the quality factor is determined in [70,95]. The second inputs in the training and testing processes are implemented using the GBVS-Enhanced algorithm.

4.3 Comparison with State-of-the-Arts

We compare the performance of Dual-ResShift with five common single image SR methods on 64 \rightarrow 256 SR tasks: BSRGAN [42], RealESRGAN [34], SwinIR [18], LDM-15 [24], and ResShift [41]. The LDM experiment is 15 steps, respectively, and the diffusion step of ResShift is 15. The evaluation metrics used are PSNR [14], SSIM [36], and LPIPS [43]. The experimental results are shown in the Table 1.

Table 1. Quantitative results of different methods on the clinical dataset and AVT dataset. The best results are highlighted in bold.

Dataset	Clinical Datasets			AVT Datasets		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
BSRGAN [42]	21.1595	0.5555	0.0876	27.3558	0.7352	0.0556
RealESRGAN [34]	24.8320	0.7214	0.0543	18.8666	0.4403	0.1307
SwinIR [18]	22.0735	0.6299	0.1590	28.1379	0.7797	0.0838
LDM-15 [24]	12.8242	0.2107	0.2745	17.7781	0.1778	0.1680
ResShift [41]	16.9386	0.4448	0.1586	23.3432	0.6704	0.0793
Dual-ResShift(Ours)	26.4165	0.7440	0.0484	28.6791	0.7805	0.0365

As seen from Table 1, Dual-ResShift is superior to other SR methods in all indicators of 64 \rightarrow 256 tasks. Specifically, Dual-ResShift exceeded Baseline(ResShift [41]) by 9.48(db) on PSNR, SSIM increased by 0.2992, and LPIPS decreased by 0.15 on the clinical dataset. On the other hand, Dual-ResShift exceeded Baseline(ResShift [41]) by 5.34(db) on PSNR, SSIM increased by 0.1101, and LPIPS decreased by 0.43 on the AVT dataset.

From Fig. 6 and Fig. 7(for ease of presentation, we enlarge the 64 * 64 LR image to 256 * 256 by bicubic.), it can be seen that the image reconstructed

by our proposed Dual-ResShift method has a richer texture and clearer edges. BSRGAN, RealESRGAN, SwinIR, and ResShift can restore basic image information, but the texture still has shortcomings. Some places are too smooth. LDM-15 fails to produce correct images, possibly due to the model's insufficient ability to handle noise in the dataset.

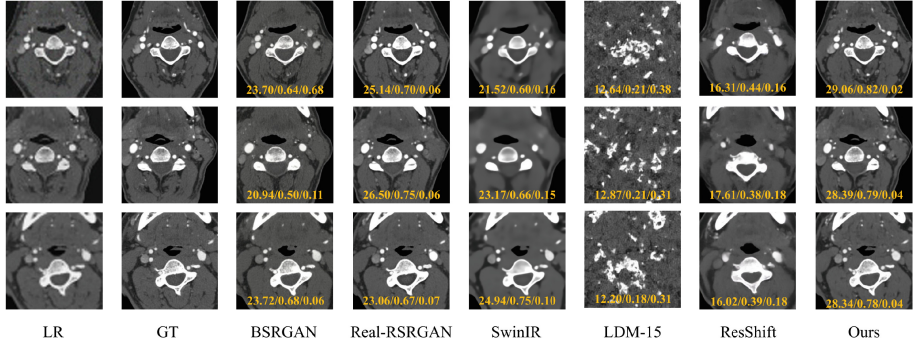


Fig. 6. Qualitative comparisons of different methods on the clinical dataset.

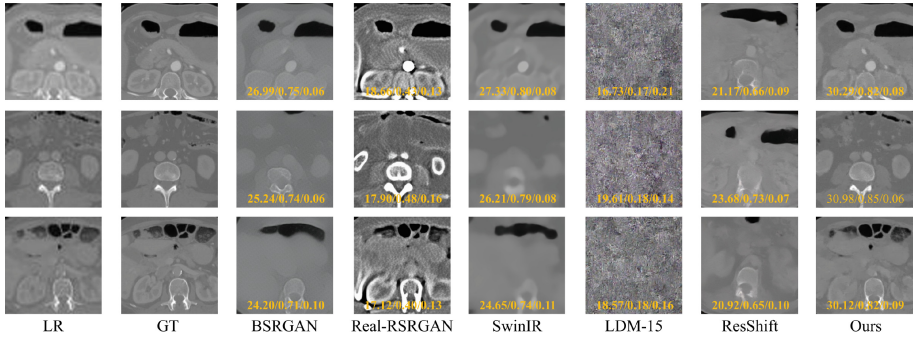


Fig. 7. Qualitative comparisons of different methods on the AVT dataset.

4.4 Ablation Study

The loss function guides the model to converge better. In order to generate higher-quality images, we evaluated the impact of the loss function in the model on the model performance. We also removed the GBVS-Enhanced algorithm as a second input to the model in the experiment to evaluate the effectiveness of the GBVS-Enhanced algorithm: 1) $\omega/o \mathcal{L}_{MSE}$ - implementing our model without \mathcal{L}_{MSE} ; 2) $\omega/o \mathcal{L}_{SF}$ - implement our model without l_{SF} ; 3) $\omega/o \mathcal{L}_{MS-SSIM}$

- implement our model without $\mathcal{L}_{MS-SSIM}$ loss function; 4) $\omega/o \mathcal{L}_{Edge}$ - implement our model without \mathcal{L}_{Edge} loss function; 5) $\omega/o GBVS - Enhanced$ - implement our model without GBVS-Enhanced algorithm. The experimental results in the clinical data set are shown in Table 2.

Table 2. Ablation study on the clinical dataset on 64 \rightarrow 256 task.

Type	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
$\omega/o \mathcal{L}_{MSE}$	25.2357	0.6924	0.0538
$\omega/o \mathcal{L}_{SF}$	25.9016	0.7158	0.0463
$\omega/o \mathcal{L}_{MS-SSIM}$	25.2770	0.6818	0.0508
$\omega/o \mathcal{L}_{Edge}$	25.5308	0.6978	0.0492
$\omega/o GBVS - Enhanced$	25.5107	0.7101	0.0491
Dual-ResShift(Ours)	26.4165	0.7440	0.0484

As can be seen from Table 2, \mathcal{L}_{Total} is slightly inferior to $\omega/o \mathcal{L}_{SF}$ in the LPIPS index by 0.021 higher by 0.5149 in PSNR and 0.0282 in SSIM. Therefore, $\omega/o \mathcal{L}_{SF}$ is helpful to the optimization of PSNR and SSIM. In other cases, \mathcal{L}_{Total} 's performance results better. Overall, \mathcal{L}_{Total} has some advantages, proving the importance of using the joint loss function in the model.

In addition, to verify the effectiveness of our proposed GBVS-Enhanced algorithm, we replaced the second input of SFUNet with the LR input. In addition, the GVBS-Ehanced algorithm is excluded without changing the structure of the model. Experiments show that, without the GBVS-Enhanced algorithm, All indexes of the model decline, so our GBVS-Enhanced algorithm helps implement SR tasks in CTA images.

5 Conclusion

We propose Dual-ResShift, a diffusion model for blind SR of CTA images. To overcome the disadvantages of previous SR methods in recovering texture and edge information, we introduce a GBVS-Enhanced algorithm to provide more information. In addition, we propose a feature-separated SFUNet to extract richer features and design a novel \mathcal{L}_{Edge} loss function to guide model optimization. Our experiments demonstrate the potential of the diffusion model for CTA SR reconstruction. In the later more in-depth research work, we will further optimize the model, consider the 3D characteristics of CTA images, enhance the impact of context on image reconstruction, and improve the quality of SR images.

Acknowledgement. This study was supported by the Fundamental Research Funds for the Central Universities(2023CDJYGRH-ZD06, 2022CDJYGRH-015), and the Chongqing medical scientific research project (Joint project of Chongqing

Health Commission and Science and Technology Bureau) (CSTB2023TIAD-KPX0050, 2024ZDXM007), and Chongqing Technology Innovation and Application Development Project(CSTB2022TIAD-KPX0176).

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
3. Chen, Y., Xie, Y., Zhou, Z., Shi, F., Christodoulou, A.G., Li, D.: Brain mri super resolution using 3d deep densely connected neural networks. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 739–742. IEEE (2018)
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
5. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Adv. Neural. Inf. Process. Syst.* **34**, 8780–8794 (2021)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part IV 13. pp. 184–199. Springer (2014)
8. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14. pp. 391–407. Springer (2016)
9. Feng, C.M., Fu, H., Yuan, S., Xu, Y.: Multi-contrast mri super-resolution via a multi-stage integration network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 140–149. Springer (2021)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020)
11. Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., Timofte, R.: Div8k: Diverse 8k resolution image dataset. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3512–3516. IEEE (2019)
12. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. *Advances in neural information processing systems* **19** (2006)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Adv. Neural. Inf. Process. Syst.* **33**, 6840–6851 (2020)
14. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. *Electron. Lett.* **44**(13), 800–801 (2008)

15. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
16. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **479**, 47–59 (2022)
17. Li, P., Li, Z., Pang, X., Wang, H., Lin, W., Wu, W.: Multi-scale residual denoising gan model for producing super-resolution cta images. *Journal of Ambient Intelligence and Humanized Computing* pp. 1–10 (2022)
18. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
19. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
20. Mao, Y., Jiang, L., Chen, X., Li, C.: Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 387–397. Springer (2023)
21. Qiu, D., Cheng, Y., Wang, X.: Residual dense attention networks for covid-19 computed tomography images super-resolution. *IEEE Transactions on Cognitive and Developmental Systems* (2022)
22. Radl, L., Jin, Y., Pepe, A., Li, J., Gsaxner, C., hua Zhao, F., Egger, J.: Aortic Vessel Tree (AVT) CTA Datasets and Segmentations (1 2022). <https://doi.org/10.6084/m9.figshare.14806362.v1>, https://figshare.com/articles/dataset/Aortic_Vessel_Tree_AVT_CTA_Datasets_and_Segmentations/14806362
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
25. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(4), 4713–4726 (2022)
26. Shao, D., Qin, L., Xiang, Y., Ma, L., Xu, H.: Medical image blind super-resolution based on improved degradation process. *IET Image Proc.* **17**(5), 1615–1625 (2023)
27. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint [arXiv:2010.02502](https://arxiv.org/abs/2010.02502) (2020)
29. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019)
30. Song, Z., Qiu, D., Zhao, X., Liu, R., Hui, Y., Jiang, H.: Parallel alternating iterative optimization for cardiac magnetic resonance image blind super-resolution. *IEEE Journal of Biomedical and Health Informatics* pp. 1–11 (2024)<https://doi.org/10.1109/JBHI.2024.3357988>

31. Wang, B., Liao, X., Ni, Y., Zhang, L., Liang, J., Wang, J., Liu, Y., Sun, X., Ou, Y., Wu, Q., et al.: High-resolution medical image reconstruction based on residual neural network for diagnosis of cerebral aneurysm. *Frontiers in Cardiovascular Medicine* **9**, 1013031 (2022)
32. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint [arXiv:2305.07015](https://arxiv.org/abs/2305.07015) (2023)
33. Wang, J., Levman, J., Pinaya, W.H.L., Tudosiu, P.D., Cardoso, M.J., Marinescu, R.: Inverser: 3d brain mri super-resolution using a latent diffusion model. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 438–447. Springer (2023)
34. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 1905–1914 (2021)
35. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3365–3387 (2020)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
37. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. vol. 2, pp. 1398–1402. Ieee (2003)
38. 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)
39. Wu, Z., Chen, X., Xie, S., Shen, J., Zeng, Y.: Super-resolution of brain mri images based on denoising diffusion probabilistic model. *Biomed. Signal Process. Control* **85**, 104901 (2023)
40. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 13095–13105 (2023)
41. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems* **36** (2024)
42. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4791–4800 (2021)
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
44. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 286–301 (2018)
45. Zhou, H., Huang, Y., Li, Y., Zhou, Y., Zheng, Y.: Blind super-resolution of 3d mri via unsupervised domain transformation. *IEEE J. Biomed. Health Inform.* **27**(3), 1409–1418 (2023). <https://doi.org/10.1109/JBHI.2022.323251>
46. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networkss. In: *Computer Vision (ICCV), 2017 IEEE International Conference on* (2017)



Multi-Block U-Net for Wind Noise Reduction in Hearing Aids

Arth J. Shah^(✉), Manish Suthar, and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar, India
{202101154,hemant_patil}@daiict.ac.in

Abstract. With advancement in technology, several changes can be observed in various technological equipments. With advancement in hearing aids technology, hearing disabled individuals are benefited abundantly. In this study, we aim to improve hearing aid technology, by proposing an advanced solution to one of the major problem of wind noise disturbance. In particular, we design a three-stage multi-block U-Net model to suppress the degradation with high quality reproduction of sound. We analyzed time-frequency domain-based audio representation analysis, and trained model on realistic noise for better user experience in hearing aids. The properties of wind noise, affecting the signal quality have been discussed deeply in addition to effect of wind noise for hearing aids users. Fully-convoluted two different blocks of U-Net were used in order to generate the proposed model, which outperforms existing models when evaluated with different performance metrics. The importance of deep learning methodologies in combination with hearing aids chips, and importance of realistic data for training an balanced model is also demonstrated in this study.

Keywords: Hearing Aids · Wind Noise Reduction · U-Net · End-to-End Denoising · Speech Enhancement

1 Introduction

After the invention in 19th century, audio recording technologies and devices have been upgrading continuously, and so is hearing aids technology. With advancement in technology, hearing aids users have been benefited abundantly by improved quality of speech, availability of different features, and functions in hearing aids. According to recently released reports of World Health Organization (WHO), 1.5 billion people (nearly 20 % of humans) have hearing loss, among whom 439 million have disabling hearing loss [1]. These 1.5 billion numbers can grow upto 2.5 billion people by the end of year 2050 as stated by reports [2]. While more than 400 million people worldwide could benefit from hearing aid use alone, only 17 % get to use these devices. An hearing disabled, individual faces many issues in communication, in particular, malfunctioning of hearing aids. One of such commonly faced issue by hearing aids users, is *Traveling Speech Degradation* (TSD), which refers to low speech quality received by hearing aids due

to noise in scene of hearing aid user. Due to such unwanted noise interruption, the quality of speech degrades abundantly, resulting into difficulty for hearing to the subject. Such noise is mainly due to high velocity of wind (i.e., wind noise), and a low frequency vehicle as well as background noise. The objective of this paper is to capture the pattern of audio waveform and decrease the background noise by employing advanced deep learning methods. Many such methods have also been explored in this field, however, they fail to capture properties of speech signal, resulting into miscasting of original speech wave.

Hearing aids users experience a variety of unwanted sounds from multiple sources, among which one of the main problem being encountered is wind noise while traveling or outstation. The quality of speech also degrades, due to *hiss* and *clicks*, resulting into low quality of speech along with wind noise. Both hiss and clicks can disrupt the clarity of the sound, making it difficult for users to focus on important sounds, such as speech. Other factors, such as listening fatigue, background noise, and distort speech sounds also make it harder to understand conversation. When this unwanted sounds mix with wind noise, the quality of speech is at its poorest version of interpretation. Moreover, it also helps us to suppress the impulsive events. Denoising of speech signals have been previously attempted using other methods, such as wavelets [3], spectral subtraction [4], linear prediction [5], and Wiener filtering [6], which were successful only for stationary noises, and resulted not giving better performance on non-stationary noise. Wind noise, is non-stationary because of its characteristics, such as amplitude and frequency, that changes over time. Wind speed and direction can vary, leading to fluctuations in the intensity and spectral properties of the noise it produces, and thereby resulting into speech signal degradation in hearing aids.

2 Related Work

In [7], the researchers employ a wind noise attenuation algorithm (WNAA) to capture properties of speech signal properties in order to decrease Signal-to-Noise Ratio (SNR) level of speech signal. They claim the wind noise to be bounded between low-mid frequency range, i.e., in regions ≤ 3 kHz. This claim may be an important clue for conducting further research on similar topic. They also illustrated and classified the effects on background noise due to direction of microphones. Study reported in [8] provides literature about directional and unidirectional microphones in behind-the-ear (BTE) hearing aids. In [9], the authors present one of the most interesting works on real-time wind noise cancellation and reduction using speech processing techniques. They employ Fast Fourier Transform (FFT)-based system, resulting into reduction of wind noise with just 32 *ms* of speech / audio frame. Also their system is known that the tolerable group delay function for mild hearing loss should be below about 5 ms. Study reported in [9] employ system that prevents spatial information for binaural hearing aids (BHA), by cancelling low delay wind noise. They employ Short-Time Fourier Transform (STFT)-based technique and estimated Perceptual Evaluation of Speech Quality (PESQ) scores for their model. Their success

motivated us to employ a similar system based on spectrogram denoising. By advancement in technology, much progress have been made in field of Machine Learning (ML), and Deep Learning (DL). Motivated by this progress, we employ a system based on advanced deep learning model, namely, U-Net, which is a convolutional-based neural network, originally proposed for image segmentation task [10].

Inspired by denoising of historic recordings [11], this study employ a system similar to them, however, the novelty lies in the structure of U-Net blocks. Study reported in [11] employed particular type of blocks (I-Blocks), which resulted in low restoration of high frequency bins while denoising. We employ multiple blocks, namely, J-Block and K-Block, in order to restore this high frequency bins while reducing effect of wind noise in speech signal. We also perform various types of analysis based on the proposed model, that validates the capability and usability of model. We employed an end-to-end system that takes noisy audio file as input, and gives output of clean denoised audio file. For robust model, we employed 6 types of different wind noise at various SNR levels. The proposed system provides the following novelty:

- Deep understanding of wind noise effect on hearing aids.
- U-Net multiblock system.
- STFT-based analysis for wind noise reduction.

The rest of the paper is organized as follows : Section 3 provides information and properties of wind noise in hearing aids. Section 4 gives details of proposed U-Net model, and methodology employed. Dataset and performance metrics detailed information are proposed in Section 5. Experimental results, their discussion, and different SNR related noise experiments are discussed in detail in Section 6. Section 7 summarizes and concludes the paper along with potential future research directions.

3 Wind Noise

This Section describes wind noise, its characteristics, and its relationship with hearing aids. Hearing aids users face a variety of acoustic environments in day-to-day life. Classification of these environments for hearing aids have also been attempted in the literature. This classification task for hearing aids is also known as Acoustic Scene Classification (ASC) [12], which is one of the primitive and basic task for speech enhancement in hearing aids. This study focus on primary and one of the most challenging task of speech enhancement in hearing aids. *Wind noise* in this study refers to air pressure creating blockage at microphones in hearing aids resulting into undesired sound at the output as hearing aids. While traveling on an vehicle (specially 2 - wheeler, i.e., bike), in deserts (where velocity of wind is much higher), in front of air conditioner (where air directly strikes hearing aids), and while running, are most common unavoidable circumstances, where wind noise in hearing aids increases abundantly and hence,

reducing speech quality of speakers. In such scenario, hearing aids user face unacceptable circumstances when the adjacent speakers' voice is affected abundantly due to high power wind noise. This results in malfunctioning of hearing aids and may cause severe fatal accidents to the users and thus, degrade quality of life. For such observations, a bunch of hearing aids users were examined in presence of wind noise, resulting into increase in hearing deficiency at that moment [7]. According to reports conducted by ORCA-US, 42 % users reported negative feedback with hearing aids in presence of wind noise [13]. Motivated by this, we propose use of an alternate model into hearing aids as an solution to this problem. While physical movement of hearing aids user, there exists two types of wind, i.e., true wind and head wind, whose vector sum results in apparent wind as mentioned in Eq. (1). In particular,

$$W_{apparent} = W_{true} + W_{head}. \quad (1)$$

For example, if a hearing aids user is running at speed of 2 m/s opposite to 2 m/s wind, apparent wind could be calculated as 4 m/s , however, if the runner and wind are in the same direction with the same velocity, apparent velocity becomes 0 m/s .

3.1 Wind Noise Creation

Low velocity of air flowing around an object, result into parallel moving (also referred to as laminar flow), and air with higher velocity cannot go around object laminar. This phenomena results in changing of direction of airflow or returning of air into the same direction as proposition generating spatial pressure difference between layers. Such partial pressure difference are referred to as turbulence, resulting into pressure variation on hearing aid microphones due to velocity variations caused by irregular airflow as shown in Fig. 1 (a). The disturbance in hearing aids output due to this partial difference is known as "wind noise" [14]. Sometimes pressure exerted by wind noise moves the diaphragm to microphone amplifier, resulting into noise distortion. Red rays on Fig. 1 (b), refers to airflow directions, which creates loud wind noise as compared to blue region resulting into quiet wind noise w.r.t. microphone position. U-Net architecture is particularly suitable for tasks like audio denoising due to its unique characteristics that facilitate efficient feature extraction and reconstruction [15].

3.2 Characteristics of Wind Noise

Studies in the literature proved that due to high force of wind noise, wind noise has high levels as 116 dB for some BTE hearing aids at low wind speed of 11 m/s , due to high force on hearing aids microphone by wind. Characteristics of the wind noise can be obtained by two major factors, namely, wind speed and wind direction. Also wind noise level is known to be proportional to square of wind speed [16, 17]. In real life measurements, as wind speed increases, the noise level can be increased much more than theoretical analysis. As discussed in

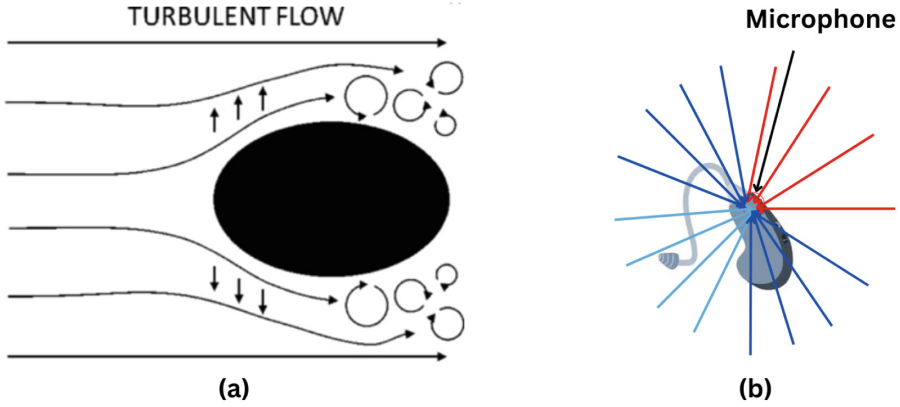


Fig. 1. (a) Turbulence created by wind [7], and (b) hearing aids noise effect w.r.t. microphone.

sub-Section 3.1, when wind noise faces directly to hearing aids microphone (red direction), the noise level is recorded to be more as compared to the wind passing through side portions (blue direction). Opposite direction (sky blue) rays state almost negligible impact of wind noise on hearing aids. Wind noise possesses many spectral characteristics as defined below :

- Low frequency energy concentration (below 300 Hz).
- Increase in spread of noise energy at high frequencies.
- Unique turbulence creation at each measurement point, resulting into rapid decrease in correlation between two points [18].
- Dual microphone hearing aids result in difference in turbulence creation at individual location.

Many mechanical changes and approaches have been explored in order to reduce wind noise effect on hearing aids, among which placing an cover over hearing aids microphone to laminate wind flow seems to have immense potential [8]. Many other approaches have also been explored on the same problem [7, 19–23].

3.3 Noise Degradation

Noise comes with various degradation in SNR levels, for this study, we choose various SNR levels including severe noisy conditions to make model more robust to different noisy environments, variable wind speed, and speech scene and thus, increase the practical suitability of our work. In moving vehicle or any other moving condition, the velocity never remains constant, and so doesn't wind speed. For such robust analysis, we selected 6 different noise conditions at different SNR levels (varying from -10 to 10 dB) in order to make model robust as much as possible.

3.4 Spectrographic Analysis

We feed STFT spectrogram as an input to U-Net model. Fig. 2 (a) represents clean audio spectrogram (STFT), and (b) shows noisy spectrogram of wind noise signal, which we aim to denoise. Fig. 2 (c), (d), (e), (f), and (g) represents noisy spectrogram, spectrogram obtained from spectral gatting, deepfilternet3, Nsnet2_denoiser, two-stage U-Net, and multistage-mutiblock U-Net of audio file. It can be observed that the spectrogram as shown in Fig. 2 (g), has more clear harmonics (black box, and white box) as compared to other obtained spectrograms. It can be also observed that in process of obtaining clean spectrogram from highly noisy spectrogram, we are able to predict the spectrogram, which resembles almost like clean spectrogram). However, high frequency region having low resolution and blunt harmonics can be observed within the highlighted area in Fig. 2 (g), indicating minor loss of a few properties while gaining the clean audio signal from noisy audio signal. Significant difference can be observed between noisy spectrogram and predicted spectrogram in other approaches (red boxes), indicating the better working of proposed model. However, training of model can be done on higher number of epoch and different variety of noise, in order to obtain a spectrogram nearly similar to clean spectrogram. After denoising, we employed standard and openly available speech enhancement model, namely, Resemble Enhancement Model (REM)¹. The audio files obtained were also less audible in other models due to loss of important data points, however, they were more efficient than the original noisy audio file as we can observe their spectrographic difference in Fig. 2.

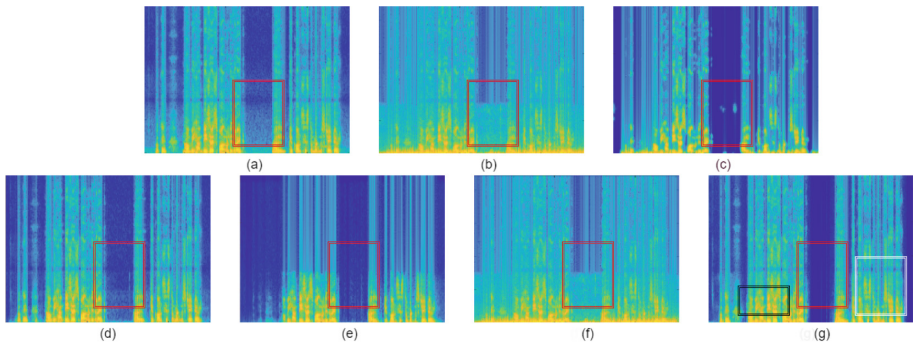


Fig. 2. (a), (b), (c), (d), (e), (f), and (g) represents clean spectrogram, noisy spectrogram, spectrogram obtained from spectral gatting, deepfilternet3, Nsnet2_denoiser, two stage unet, and multistage-mutiblock unet of audio file, respectively.

¹ resemble-ai {Last Accessed Date : 29th July, 2024}

4 Proposed Approach

This Section describes the model (spectrogram-based fully-convoluted) employed in this study in order to minimize the impact of wind noise on the speech signal. Much progress have been made in the field of image restoration and denoising in recent days, among whom study reported in [24] demonstrates a multistage image restoration architecture of U-Net. To that effect, we employ a supervised attention module (SAM)-based three-stage multiblock U-Net architecture for wind noise reduction, as shown in Fig. 4. Two phase approach has also been motivated by an recent study [11], which also employs almost similar architecture for historical recordings restoration. In the first phase of two phase method consists of U-Net sub-networks. However, on the first stage, the inputs and training objective differs. The aim of this stage is to identify and localize the residual noise emerging after denoising a speech signal. The second stage works as an main classifier, which also ensures smooth denoising, and refines the noisy speech using extracted features from first stage. This will decrease the impact of wind noise over speech signal. We decided to explore Short-Time Fourier Transform (STFT) spectrogram as input to U-Net module. STFT is useful when analyzing non-stationary signal having different spectral content over time. It involves segmenting an audio signal into small time intervals and performing Fourier transform (FT) on those segments. Unlike the standard FT, which provides a global frequency view (because of infinite support of Fourier integral), the STFT captures local frequency content that varies with time. Each FT in the STFT provides simultaneous time and frequency information. It highlights how different frequencies contribute to the signal over short time windows. This is particularly useful for non-stationary signals, such as, speech and music, where frequency content changes dynamically. Inorder to capture dynamic change in speech signal, and local frequency components, we extracted STFT spectrograms from 44.1 kHz audio signal, thereafter reading real and imaginary part of signal as real-valued signal. Frame size and frame length were chosen as 2048 samples, and 512 samples, respectively. In the initial layers, we enhance the network’s frequency information by including frequency-positional embedding [11] as 10 additional channels in the input data. In each phase, the 12-channel input coming off of the previous process is passed through a shallow feature extractor, comprised of a convolutional layer followed by the Exponential Linear Unit (ELU) [25] as shown in Fig. 4. $F_{in,1}$ extracted at the first layer is then fed directly into U-Net sub-network; whereas for the second phase, we concatenate input features ($F_{in,2}$) with the ones from a set of other features produced by an additive component in the attention mask for each position. Only optimal features emerges from second stage of neural network, due to SAM module. The U-Net output features $F_{out,1}$ are used to create the estimated residual noise signal N via a 3×3 convolutional layer. The first stage output $Y1$ is estimated as $Y1 = X + N$, where X is the input spectrumgram. The attention-guided features F_{SAM} presented in Fig. 4 are determined by the attention masks M produced from $Y1$ through a 1×1 convolution, and a sigmoid function. Lastly, we take the features output from

the second U-Net ($F_{out,2}$) then apply a 3×3 conv layer to generate the denoised output, Y_2 .

To supervise the model, we minimize the mean absolute error of both outputs at each stage. The reconstruction loss function is defined as:

$$L = \frac{1}{K} \sum_{k=1}^K (|Y_{k1} - Y_k| + |Y_{k2} - Y_k|). \quad (2)$$

Consider a case where clean spectrogram is represented by Y , and the total number of STFT bins is K . The Adam optimizer [30] with $\beta_1 = 0.5$, $\beta_2 = 0.9$, and a starting learning rate of 1×10^{-4} decaying by a factor of 10 for every 100,000 steps was used during training. It is worth noting that normalization strategies were not used since batch normalization and weight normalization did not produce any improvements in our experiments.

4.1 U-Net Subnetworks

In computer vision and audio processing, the most common use of the U-Net architecture is documented [26], [27]. We employed U-Net sub-networks that are designed following a symmetrical encoder-decoder design, featuring four downsamplers and upsamplers as shown in Fig. 3 (a) for this study. Each scale contains an intermediate block referred to as either J-Block 3 (b), I-Block 3 (c) or K-Block 3 (d) as shown in Fig. 3, which consists of DenseNet two-layer block with residual connection. Concatenating skip connections are used to connect the outputs of the encoder J-Blocks to their respective decoder J-Blocks. Additionally, an I-Block is placed after the fourth downsampler. K-Block is similar to J-Block, however, the number of DenseNet block with residual connection increases to 4. The downsampling layers employ strided convolutions with a stride of 2×2 and a kernel size of 4×4 as well as the same number of filters as the next I-Block, which consists of three DenseNet blocks, as illustrated in Fig. 3 (c). In the decoder, the upsampling layers make use of transposed convolutions having identical hyperparameters to their corresponding downsampling layers. Our experiments showed that while checkerboard artifacts are a common side effect of transposed convolutions, they started to disappear as training progressed.

4.2 Resemble Enhancement Model

The Resemble Enhancement Model (REM)¹ is the ultimate in sound enhancement technology. REM is based on sophisticated machine learning algorithms, which allows it to do more than just analyse sound. It enhances speech intelligibility, and maintains the original tonality of music and voice. Being a seasoned sound designer, it naturally modifies his techniques to fit various settings and sound kinds. REM works excellent for enhancing speaker volume in a conference context. Its true allure is in its capacity to isolate and accentuate specific audio components while preserving a clean, distortion-free sound quality efficient fine-tuning algorithms that guarantee minimum latency and maximum effect are

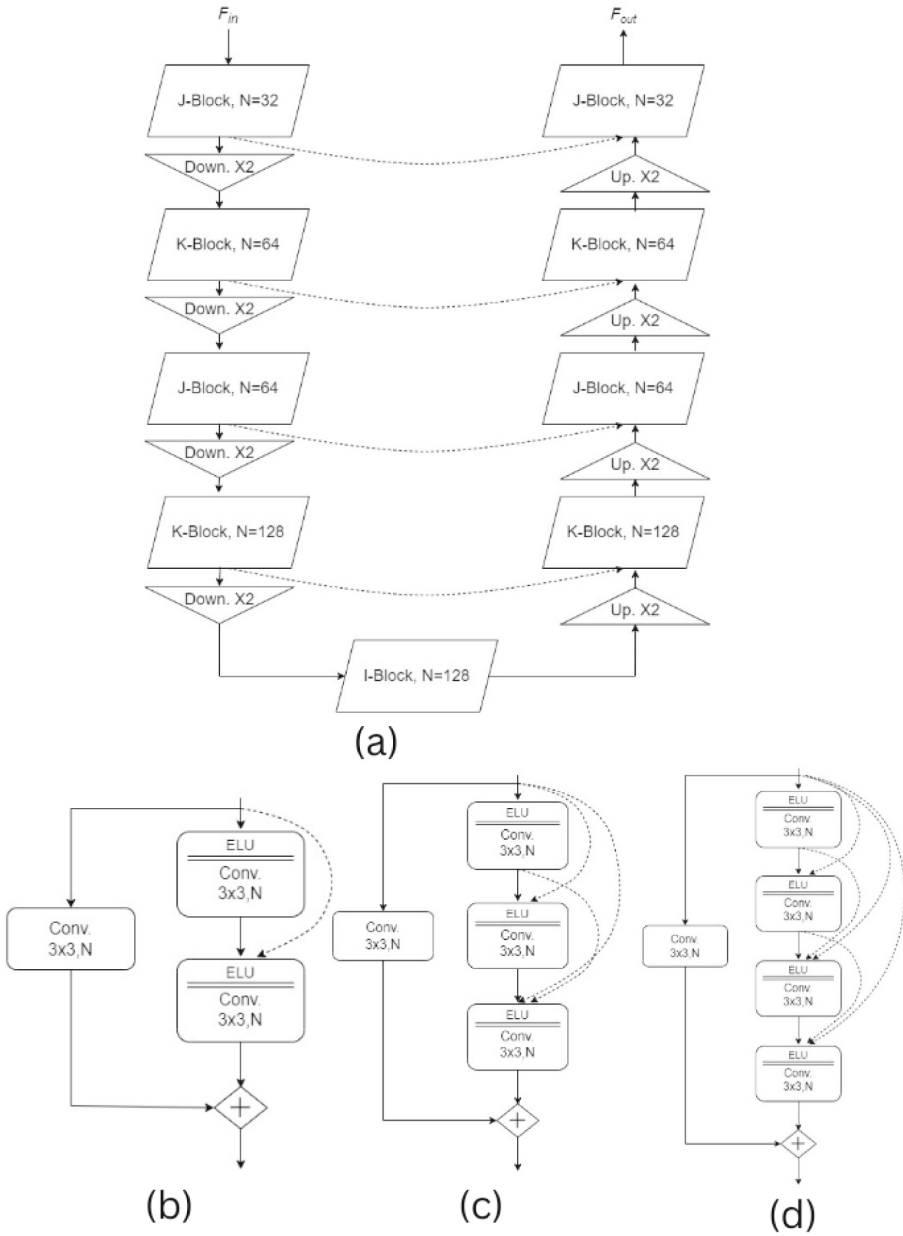


Fig. 3. SAM module embedded with UNet architecture, and blocks After [11].

responsible for this quick performance. The enhancer is a latent conditional flow matching (CFM) model. It consists of an Implicit Rank-Minimizing Autoencoder (IRMAE), and a CFM model that predicts the latents. This first stage involves

an autoencoder that compresses the clean Mel spectrogram (M) clean into a compact latent representation (Z) clean , which is then decoded and vocoded back into a waveform. The model consists of an encoder, decoder, and vocoder.

As we have used pre-trained model, it works as follows. After completing the training of the first stage, in second stage the latent CFM model is trained. The CFM model is conditioned on a blended Mel, M blend = αM denoised + $(1 - \alpha)M$ noisy , derived from the noisy STFT-spectrogram M noisy and a denoised STFT-spectrogram M denoised Here, α is the parameter that adjusts the strength of the denoiser. During training, α is set to follow a uniform distribution $U(0,1)$. During inference, value of α can be controlled by the user. To predict the latent representation of the clean speech, we have used the loaded pre-trained enhancer model and which is already trained jointly with the latent CFM model. REM does not discriminate between sources and audio formats. REM effortlessly transitions between clear speech recordings, energetic musical performances, and tranquil surround sound to provide an improved listening experience on all media platforms and playback devices. Using this three-stage U-Net, we are able to obtain enhanced speech in less than 2 seconds, which in future we aim to reduce.

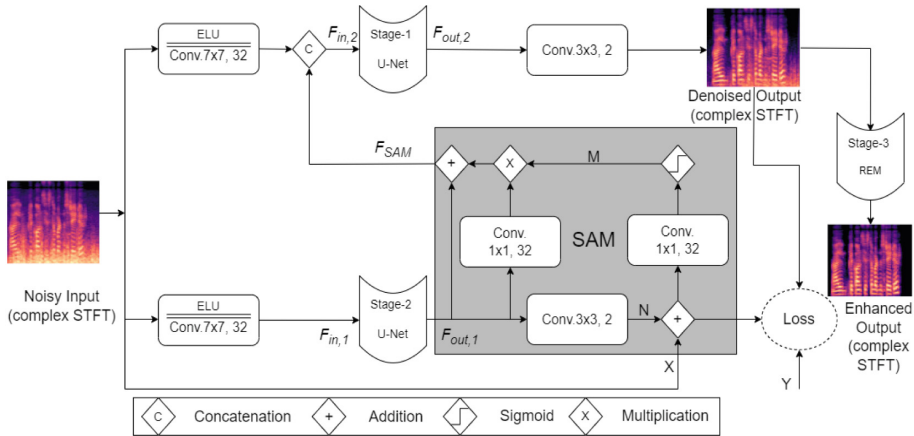


Fig. 4. Architecture of three stage denoising model with the SAM module. After [11].

5 Experimental Setup

5.1 Dataset Used

The CHiME-1 (Computational Hearing in Multisource Environments) dataset is the first installment in a series designed to facilitate the development and evaluation of robust speech recognition systems in challenging acoustic environments.

Recorded in a typical domestic setting, the dataset features binaural recordings of utterances from the GRID corpus, a collection of simple, fixed grammar sentences, spoken by multiple speakers. The CHiME-1 dataset includes both clean and noisy versions of the recordings, allowing us to do controlled experimentation in noise robustness. Additionally, it provides detailed transcriptions of the speech content and separate recordings of the background noises, enabling researchers to test and improve noise suppression and speech enhancement algorithms. This dataset has been widely used for benchmarking and advancing techniques in the field of far-field speech recognition, particularly in scenarios involving multiple, and dynamic noise sources. Six types of different wind noise were collected from freesound².

5.2 Performance Metrics Used

Mean Absolute Error (MAE) : It measures the average magnitude of errors between predicted and actual values without considering their direction. It is calculated as the average of the absolute differences between predictions and actual observations, providing a straightforward interpretation of the prediction accuracy.

Coherence : In the context of topic modeling, coherence measures the degree of semantic similarity between high-scoring words in a topic. A higher coherence score indicates that the words within a topic are more related to each other, suggesting better quality of the topic model.

Δ Mean Squared Deviation (Δ MSD) : It is a variation of Mean Squared Deviation (MSD), often used to assess the variability of a set of values. It measures the squared differences between predicted and actual values, providing a sense of how predictions deviate from actual outcomes. Delta MSD specifically emphasizes the changes or differences between these deviations over time or across different conditions.

Short-Time Objective Intelligibility (STOI) : STOI is an objective metric for evaluating speech intelligibility. It compares short-time segments of the clean and degraded speech signals to estimate how intelligible the speech is to human listeners. STOI is widely used in speech enhancement and hearing aid algorithms to assess and improve the clarity of speech signals.

6 Experimental Results

We trained model for 200 epochs, and 2000 steps per epoch. At the end of training, we were able to obtain training loss of 3 % and validation loss of around 6 %.

² freesound.org {Last Accessed Date : 29th July, 2024}.

Table 1. Comparison of proposed approach with existing works

Noise	Model	Coherence	Δ MSD	STOI	MAE
-5dB	Spectral-Gatting [28]	0.88227	5.3702	0.77388	1559.27
	DeepFilterNet [29]	0.8538	1.0882	0.9128	728.71
	Nsnet2-denoiser [30]	0.343	1.6175	0.1715	56299977.16
	Two-Stage Unet [11]	0.8972	4.9898	0.82	3854.66
	Multiblock-Unet (Without REM)	0.9091	4.4575	0.82166	3578.19
	Multistage-MultiBlock-Unet (Proposed)	0.9642	2.4395	0.9189	1037.99
0dB	Spectral-Gatting [28]	0.89179	5.5726	0.8473	1559.21
	DeepFilterNet [29]	0.97794	0.6378	0.9416	454.64
	Nsnet2-denoiser [30]	0.36411	1.2707	0.1328	41908510.01
	Two-Stage Unet [11]	0.95397	2.5139	0.8776	2108.84
	Multiblock-Unet (Without REM)	0.95741	2.2291	0.8799	1946.9
	Multistage-MultiBlock-Unet (Proposed)	0.983	2.2743	0.9516	726.57
5dB	Spectral-Gatting [28]	0.89889	5.5219	0.8658	1559.27
	DeepFilterNet [29]	0.9826	0.5256	0.9571	382.57
	Nsnet2-denoiser [30]	0.36342	0.8017	0.1627	35356174.96
	Two-Stage Unet [11]	0.97894	0.8662	0.9071	1217.19
	Multiblock-Unet (Without REM)	0.97863	0.78482	0.9094	1103.75
	Multistage-MultiBlock-Unet (Proposed)	0.98713	2.3324	0.9611	665.61
10dB	Spectral-Gatting [28]	0.90091	5.3941	0.914	1559.27
	DeepFilterNet [29]	0.98662	0.248	0.977	237.07
	Nsnet2-denoiser [30]	0.3661	0.5275	0.1313	26015898.19
	Two-Stage Unet [11]	0.98689	0.3694	0.9614	613.36
	Multiblock-Unet (Without REM)	0.98722	0.3004	0.9633	572.68
	Multistage-MultiBlock-Unet (Proposed)	0.9903	2.6847	0.9817	641.1

6.1 Comparison With Existing Works

This sub-Section compared results with few existing works. However, we were not able to compare the work with more existing works, as not all the studies released their trained model. Due to lack of time, and insufficiency of resources,

we were not able to retrain existing models, and hence, decided to compare them with open source released models available. Table 1 denotes the obtained results on various existing studies, compared with proposed methodology. Observations based on Table 1 denotes that, after completing process of multi stage U-Net on a noisy audio, the words are clear enough for a subject to be interpreted and have high coherence than the other existing models as observed. Coherence denotes that how much words within topic are related to each other. On the other hand, low coherence on existing models is due to less denoising of model, or removal of important speech properties while denoising. Alternatively, DeepFilterNet, being an advance speech denoising / enhancement model, outperforms proposed method in various aspects. At the point, controversy arises that is the proposed model an optimal model ? Resulting into an positive aspect, we also calculate Mean Opinion Score (MOS) for DeepFilterNet, resulting into proving superiority of proposed methodology. MOS being an evaluation metrics to validate model on real-life situation, by asking an subject to rate the denoised speech between 1 (Bad) to 5 (Good). Table 2 denotes the ratings of 108 users after hearing speech from various speech models. In Table 1, proposed model did not performed well on other metrics as compared to DeepFilterNet, as DeepFilterNet is denoising more as compared to proposed method, however, it also degrades the speech quality while denoising and user is not able to get better experience due to degraded quality of sound, which can be observed and concluded from results shown in Table 2.

Table 2. MOS obtained on 108 participants (78 Male + 30 Female)

Model	MOS
Spectral Gating [28]	3.67
Nsnet2-denoiser [30]	2.33
Two-Stage Unet [11]	3.89
DeepFilterNet [29]	4.21
Proposed	4.52

7 Summary and Conclusions

In this work, we presented significance of multi-stage U-Net formed from various types of blocks to denoise hearing aids speech with additive white noise. We propose end-to-end methodology for identifying the pattern of wind noise in the speech signal. This study investigated three-stage, namely, UNet, formed by three different types of blocks, namely, I-Block, J-Block, and K-Block. The primary goal of this study is to restore and enhance the quality of speech, which has been degraded due to presence of various types of wind noise at various SNR

levels. The features extracted (i.e., STFT-based spectrograms) by signal processing concepts were then fed into the U-Net model for the process of mapping a noisy spectrogram to clean spectrogram. For this study, we made a variety of observations based on models capability, such as, evaluations based on coherence, Δ MSD, STOI, MAE, and MOS. In comparison to existing widely used and open source denoising models, we achieved significantly better results for the task selected. We also discussed the difference and improvement between proposed and existing methodology. The proposed system have been explored for only six type of noise, which we aim to extend the work to various different types of noise, to analyze effect of different types of noise on model, as a future task. Future works also involve more detailed mathematics on proposed methodology, and exploring variety of blocks for proposed approach.

Acknowledgements. This study has been funded by Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the sponsorship of project 'BHASHINI', (Grant ID: 11(1)2022-HCC(TDIL)). Authors thank authorities of DA-IICT Gandhinagar for their kind support and cooperation to carry out this research work.

References

1. S. K. Swain, "Hearing loss and its impact in the community," *Matrix Science Medica*, vol. 8, no. 1, pp. 1-5, 2024, Last Accessed Date : 2ndJuly, 2024
2. W. H. Organization et al., *World report on hearing*. World Health Organization, 2021, Last Accessed Date : 5thJune, 2024
3. C. Schremmer, T. Haenselmann, and F. Bomers, "A wavelet based audio denoiser," in *Proc. IEEE International Conference on Multimedia and Expo, (ICME)*, Barcelona, Spain, 2001, 145-148, Tokyo, Japan
4. T. Biswas, C. Pal, S. B. Mandal, and A. Chakrabarti, "Audio de-noising by spectral subtraction technique implemented on reconfigurable hardware," in *2014 Seventh International Conference on Contemporary Computing (IC3)*, 2014, 236-241, Noida, India
5. Delcroix, M., Hikichi, T., Miyoshi, M.: Dereverberation and denoising using multi-channel linear prediction. *IEEE Trans. Audio Speech Lang. Process.* **15**(6), 1791–1801 (2007)
6. Xia, B., Bao, C.: Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Commun.* **60**, 13–29 (2014)
7. P. Korhonen, "Wind noise management in hearing aids," in *Seminars in Hearing*, Thieme Medical Publishers, Inc., vol. 42, 2021, pp. 248-259
8. Chung, K., Mongeau, L., McKibben, N.: Wind noise in hearing aids with directional and omnidirectional microphones: Polar characteristics of behind-the-ear hearing aids. *The Journal of the Acoustical Society of America (JASA)* **125**(4), 2243–2259 (2009)
9. Uemura, Y., Nakashima, H., Hiruma, N., Fujisaka, Y.-I.: "Real-time wind noise cancellation based on binaural cues for hearing aids," *Western Pacific Acoustics Conference (WESPAC)*. New Delhi, India (2018)

10. 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
11. Moliner, E., Välimäki, V.: A two-stage U-Net for high-fidelity denoising of historical recordings. In: International, I.E.E.E. (ed.) Conference on Acoustics, pp. 841–845. Singapore, Speech and Signal Processing (ICASSP) (2022)
12. Hüwel, A., Adiloğlu, K., Bach, J.-H.: Hearing aid research data set for acoustic environment recognition. In: International, I.E.E.E. (ed.) Conference on Acoustics, pp. 706–710. Barcelona, Speech and Signal Processing (ICASSP) (2020)
13. Kochkin, S.: Marketrak viii: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal* **63**(1), 19–20 (2010)
14. Walker, K.T., Hedlin, M.A.: A review of Wind-Noise Reduction Methodologies. *Infrasound Monitoring for Atmospheric Studies* **1**, 141–182 (2009)
15. Zakis, J.A.: Wind noise at microphones within and across hearing aids at wind speeds below and above microphone saturation. *The Journal of the Acoustical Society of America (JASA)* **129**(6), 3897–3907 (2011)
16. Strasberg, M.: Dimensional Analysis of Windscreen Noise. *The Journal of the Acoustical Society of America (JASA)* **83**(2), 544–548 (1988)
17. J. M. Kates, *Digital Hearing Aids*. Plural publishing, 2008, <https://www.pluralpublishing.com/publications/digitalhearingaids> Last Accessed Date : 10th April, 2024
18. Li, W., Liu, H.: Two-point statistics of coherent structures in turbulent flow over riblet-mounted surfaces. *Acta. Mech. Sin.* **35**(3), 457–471 (2019). <https://doi.org/10.1007/s10409-018-0828-2>
19. Widex, “Widex supertm power to hear,” <https://www.widex.biz/axapta/documents/9>
20. Ricketts, T., Dittberner, A., Johnson, E.: High-frequency amplification and sound quality in listeners with normal through moderate hearing loss. *J. Speech Lang. Hear. Res.* **51**(1), 160–172 (2008)
21. Chung, K., McKibben, N., Mongeau, L.: Wind noise in hearing aids with directional and omnidirectional microphones: Polar characteristics of custom-made hearing aids. *The Journal of the Acoustical Society of America (JASA)* **127**(4), 2529–2542 (2010)
22. G. J., “An innovative rie with microphone in the ear lets users hear with their own ears,” GN Hearing AS, pp. 1-8, 2020
23. Chung, K.: Effects of venting on wind noise levels measured at the eardrum. *Ear Hear.* **34**(4), 470–481 (2013)
24. S. W. Zamir, A. Arora, S. Khan, et al., “Multi-stage progressive image restoration,” in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), 2021, Virtual, pp. 14 821-14 831
25. A. Shah, E. Kadam, H. Shah, S. Shinde, and S. Shingade, “Deep residual networks with exponential linear unit,” in Proceedings of the 3rd International Symposium on Computer Vision and the Internet, 2016, Jaipur, India, pp. 59-65
26. V. Iglovikov and A. Shvets, “Ternausnet: U-net with vgg11 encoder pretrained on imagenet for image segmentation,” arXiv preprint [arXiv:1801.05746](https://arxiv.org/abs/1801.05746), 2018, Last Accessed Date : 5th June, 2024
27. Baloch, D., Abdullah, S., Qaiser, A., Ahmed, S., Nasim, F., Kanwal, M.: Speech enhancement using fully convolutional unet and gated convolutional neural network. *Int. J. Adv. Comput. Sci. Appl.* **14**, 831–836 (2023)

28. E. Sudheer Kumar, K. Jai Surya, K. Yaswanth Varma, A. Akash, and K. Nithish Reddy, "Noise reduction in audio file using spectral gatting and fft by python modules," in Recent Developments in Electronics and Communication Systems, IOS Press, 2023, pp. 510-515
29. H. Schrüter, T. Rosenkranz, A. Maier, et al., "Deepfilternet: Perceptually motivated real-time speech enhancement," arXiv preprint [arXiv:2305.08227](https://arxiv.org/abs/2305.08227), 2023, Last Accessed Date : 5thJune, 2024
30. S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in International Conference on Speech and Computer, Springer, 2020, pp. 79-86



A Cascading Approach with Vision Transformers for Age-Related Macular Degeneration Diagnosis and Explainability

Ainhoa Osa-Sanchez^{1,2}, Hossam Magdy Balaha^{1(✉)}, Mahmoud Ali¹, Mostafa Abdelrahim¹, Mohmaed Khudri¹, Begonya Garcia-Zapirain², and Ayman El-Baz¹

¹ Department of Bioengineering, J.B. Speed School of Engineering, University of Louisville, Louisville, KY, USA
hmbala01@louisville.edu

² eVIDA Research Group, University of Deusto, 48007 Bilbao, Spain

Abstract. Age-related macular degeneration (AMD) progressively damages the macula, the central area of the retina crucial for sharp vision. While early and intermediate stages may be asymptomatic, advanced AMD can lead to significant vision loss, affecting tasks such as reading and facial recognition. The proposed framework employs a cascading approach with two integrated stages for AMD diagnosis, encompassing data preprocessing, model training, and cascaded prediction with augmentation and tuning. Utilizing Vision Transformers (ViT), renowned for their ability to handle intricate image features via self-attention mechanisms, the framework integrates three distinct ViT classifiers (\mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3). Each classifier specializes in differentiating AMD conditions based on patient data and image characteristics. The cascade model iteratively refines predictions across these stages, ensuring robust diagnostic accuracy tailored to diverse AMD conditions. Interpretability is enhanced using SHAP, LIME, and GradCAM techniques, providing insights into model decision-making and validating automated diagnoses within retinal imaging for AMD. The proposed cascaded approach achieves an accuracy of 93.18%, recall of 94.44%, and specificity of 91.18%.

Keywords: Age-related Macular Degeneration (AMD) · eXplainable Artificial Intelligence (XAI) · Vision Transformer (ViT) · SHapley Additive exPlanations (SHAP) · Local Interpretable Model-agnostic Explanations (LIME) · Gradient-weighted Class Activation Mapping (GradCAM)

1 Introduction

Age-related macular degeneration (AMD) progressively damages the macula, the vital central area of the retina responsible for sharp vision. Although early and intermediate stages of AMD may go unnoticed, advanced AMD can result in

A. Osa-Sanchez and H. M. Balaha—These authors contributed equally to this work

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15327, pp. 250–265, 2025.

https://doi.org/10.1007/978-3-031-78398-2_17

severe vision impairment, impacting activities like reading and recognizing faces [2, 10, 23]. AMD is a leading cause of vision loss worldwide among those aged 55 and older, accounting for 6% to 9% of global cases of legal blindness across all income levels. Projections suggest a significant rise in affected individuals globally, from 196 million in 2020 to 288 million by 2040 [3, 4, 8]. In the United States, approximately 20 million people were living with AMD in 2019, with about 1.5 million of them experiencing advanced AMD.

AMD comes in two forms: dry and wet, each with its stages. Dry AMD progresses through early, intermediate, and late stages, the latter termed geographic atrophy (or advanced non-neovascular AMD). Conversely, wet AMD, also known as exudative or neovascular AMD, mainly occurs in the late stage, is characterized by inactive and active grades, and is further classified as classic, occult, or mixed. Both wet AMD and late-stage dry AMD are advanced forms of the disease [7]. Although dry AMD is more widespread, wet AMD, despite being less common, accounts for 90% of AMD-related blindness [21].

Several studies such as [12, 26] have worked on the AMD diagnosis using state-of-the-art approaches such as Vision Transformers (ViTs). ViTs have revolutionized computer vision, particularly in medical imaging diagnosis. Their advanced architecture especially in the attention mechanism and superior performance have enabled precise detection and diagnosis of retinal diseases like AMD, glaucoma, choroidal neovascularization (CNV), and diabetic macular edema (DME) [5, 22].

ViT excels across various imaging modalities, including retinal fundus images, optical coherence tomography (OCT), and Spectral Domain Optical Coherence Tomography (SD-OCT). Utilizing transformer-based architecture, ViT captures intricate patterns in retinal images, significantly enhancing diagnostic accuracy. As a result, ViT has become indispensable in medical imaging [19].

The current study suggests a computer-aided diagnosis (CAD) system that is capable of accurately distinguishing between normal retinas (non-AMD), intermediate dry AMD, GA, and wet AMD grades using fundus images, while effectively addressing the variability in fundus image dimensions. It employs the state-of-the-art ViT for diagnosis and eXplainable AI (XAI) for interpretability. Medical doctors subsequently validate the XAI results. The diagnostic process occurs in two cascading phases. In the first phase, the system diagnoses whether the patient has Geographic Atrophy (GA)/Intermediate or Normal/Wet. The second phase conducts further diagnosis based on the output from the first phase. Overall, there are three ViT classifiers involved in the process.

2 Related Studies

Recent studies have tackled the problem of AMD diagnosis. For instance, Ghomami et al. [9] investigated federated learning (FL) for diagnosing AMD using deep learning (DL) models. They aimed to enhance diagnostic accuracy while preserving data privacy. Centralized models, particularly with the ViT encoder, demonstrated exceptional performance, with accuracy rates of $98.18\% \pm 0.55$ and $99.11\% \pm 0.39$ on Dataset 2 and 3 test sets, respectively.

In addition, Xu et al. [26] introduced DeepDrAMD, an Artificial Intelligence (AI) model for automated AMD detection and subtype classification, achieving high Area Under the Curves (AUCs) of 98.76% and 96.47% in different test cohorts. Furthermore, Yao et al. [27] presented the FunSwin, a method for grading diabetic retinopathy and estimating macular edema risk, outperforming existing studies with an accuracy of 98.66% and an F1 score of 98.96% for binary classification of macular edema.

Additionally, Kihara et al. [13] developed a DL model utilizing a ViT architecture to detect nonexudative macular neovascularization (neMNV) from OCT B-scans. Using Swept-Source Optical Coherence Tomography Angiography (SS-OCTA) imaging, B-scans were annotated to distinguish between Drusen and neMNV-associated double-layer sign (DLS). The ViT model's performance was compared to human graders, demonstrating 82% sensitivity, 90% specificity, and strong agreement with senior human graders ($\kappa = 0.83$, $P < 0.001$).

Moreover, Akcca et al. [5] introduced 3 algorithms, ViT, Tokens-To-Token Vision Transformer (T2T-ViT), and Mobile ViT, aimed at detecting CNV, drusen, DME, and normal features within OCT images. ViT, T2T-ViT, and Mobile-ViT achieved predictive accuracies of 95.14%, 96.07%, and 99.17%, respectively. Particularly noteworthy is the superior performance of Mobile-ViT, which exhibited a higher classification accuracy compared to the other models.

Besides, Jiang et al. [12] introduced a CAD method using a ViT to analyze OCT images, achieving an exceptional classification accuracy of 99.69% in distinguishing AMD, DME, and normal eyes. After model pruning, recognition time was reduced to an impressive 0.010 seconds without compromising accuracy. Compared to traditional CNN models like VGG16, ResNet50, DenseNet121, and EfficientNet, the pruned ViT demonstrated superior recognition capabilities.

Despite the impressive performance in these studies, many did not use state-of-the-art (SOTA) approaches like ViTs, and some did not incorporate explainability into their models. Therefore, in this study, we aim to address the challenges of employing ViTs and interpreting the model using explainable AI.

3 Materials

This study endeavors to address the classification challenge of discerning AMD grades by categorizing colored fundus images of patients into normal or exhibiting intermediate AMD, GA, or wet AMD grades. The approach is implemented on a localized dataset.

For this study, we utilized a dataset comprising 864 human subjects obtained through the Comparisons of AMD Treatments Trials (CATT) [1], sponsored by the University of Pennsylvania. Subjects, aged 50 and above, were enrolled over two years across 43 clinical centers in the United States. They underwent intravitreal injections of either ranibizumab or bevacizumab according to one of three dosing regimens.

Treatment administration began at the participants' initial visit under the CATT program. Participants in fixed monthly dosing groups received treatment at each visit, while those in variable dosing groups received treatment based on the presence of exudation. Evaluations occurred at every visit, with participants demonstrating lesion activity receiving treatment.

All imaging and clinical data underwent de-identification by the CATT Study Group before transmission to the University of Louisville (UofL). As the dataset was previously collected and appropriately de-identified by a third party, the study received an exemption from the local institutional review board (IRB) process at UofL. Data collection procedures adhered strictly to relevant guidelines and regulations, with informed consent obtained from all participants or their legal guardians.

The dataset comprised 216 normal, 216 intermediate AMD, 216 GA AMD, and 216 Wet AMD cases, all collected from this cohort. Samples from the dataset are presented in Figure 1.

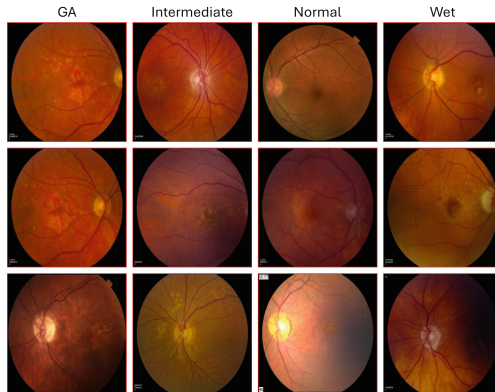


Fig. 1. Samples from the utilized AMD dataset for the four categories.

4 Methodology

The proposed framework (Figure 2) utilizes a cascading approach consisting of two integrated stages for AMD diagnosis. The study encompasses three major stages: data preprocessing, model training, and cascaded prediction with and without data augmentation. The framework employs ViT models renowned for their ability to handle complex image features through self-attention mechanisms. The two integrated stages include three distinct ViT classifiers (\mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3), each specializing in differentiating between AMD conditions based on patient data and image characteristics. The first phase determines whether the patient has GA/Intermediate or Normal/Wet conditions, and the second phase splits the diagnosis based on the outcome of the first phase. The framework's interpretability is

enhanced using SHAP, LIME, and GradCAM techniques, which provide insights into model decision-making processes and aid in validating automated diagnoses within the context of retinal imaging for AMD.

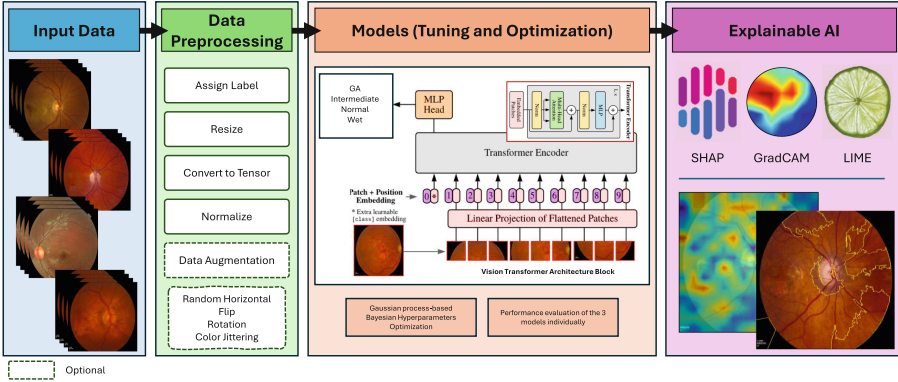


Fig. 2. The proposed framework for AMD diagnosis, tuning, and explainability.

4.1 Preprocessing

The dataset undergoes preprocessing before the classification phase, which includes resizing, normalization, and augmentation. The images are resized to 224×224 pixels and normalized with a mean of $\mu = [0.485, 0.456, 0.406]$ and a standard deviation of $\sigma = [0.229, 0.224, 0.225]$. Two scenarios were utilized to study the effect of data augmentation. In the first scenario, we conducted standard preprocessing without augmentation. This involved resizing images, converting them to tensors, and applying normalization. In contrast, the second scenario incorporated data augmentation during model training. Augmentation techniques included random horizontal flipping, random rotation (up to 10°), and color jittering (e.g., brightness, contrast, saturation, and hue adjustments).

4.2 Classification Architecture and Tuning

The diagnostic process described employs a sophisticated approach using ViT, specialized deep learning models originally designed for image classification tasks. Unlike traditional convolutional neural networks (CNNs), ViTs process images by dividing them into patches and employing self-attention mechanisms to capture global dependencies across these patches [17]. This method allows ViTs to effectively model spatial relationships within medical images, making them highly suitable for tasks where understanding global context is crucial, such as medical diagnostics [20].

In this specific diagnostic setup, the process unfolds in two cascading phases facilitated by three distinct ViT classifiers as presented in Figure 3. A cascade

model is an advanced ensemble learning approach where multiple models are employed sequentially to refine predictions. This method utilizes the strengths of various models to improve the overall accuracy and reliability of predictions [28]. Sequential execution involves a series of models, each making a prediction and moving to the next model if a predefined confidence level is achieved, continuing this process until a final decision is made or the cascade ends [14].

The first phase, handled by \mathcal{M}_1 , takes as input various features extracted from patient data. It operates as the initial classifier, categorizing patients into one of two broad groups: those exhibiting signs of GA or Intermediate conditions, and those showing signs of Normal or Wet conditions.

Following the initial classification in Phase 1, the diagnostic pathway diverges based on the \mathcal{M}_1 's output. For patients categorized as having GA or Intermediate conditions, \mathcal{M}_2 is activated in the second phase. \mathcal{M}_2 specializes in providing detailed and specific diagnoses tailored to these conditions, utilizing its capability to analyze finer details and patterns within medical images that indicate GA or Intermediate stages.

Conversely, for patients identified in Phase 1 as having Normal or Wet conditions, \mathcal{M}_3 is engaged in Phase 2. \mathcal{M}_3 is designed to offer detailed diagnoses specific to Normal or Wet conditions, focusing on identifying key markers such as fluid accumulations or vascular irregularities that characterize these conditions.

The decision logic depends on the current model's confidence level; if the prediction is uncertain or below the threshold, the next model in the sequence is used to refine or improve the prediction. Throughout the cascade, performance metrics such as accuracy, precision, recall, F1 score, and confusion matrix elements are gathered [15]. These metrics are used to evaluate the cascade model's effectiveness, with the final prediction being an aggregate of these metrics, ensuring a thorough assessment of the model's performance.

Additionally, We developed a systematic approach to train and optimize using Gaussian process-based Bayesian optimization. The architecture hyperparameters are dimensionality (dim), depth of transformer encoder layers (D), number of attention heads (H), MLP dimension (dim_{MLP}), dropout rates (α_{common} and α_{emb}), and learning rate (γ). dim determines the dimensionality of the embedding space, allowing for a more complex representation of the input data while balancing the risk of overfitting. D specifies the number of layers in the model, with greater depth enabling the modeling of more complex functions at the cost of increased computational demands and potential overfitting.

Moreover, H represents the number of attention heads in the multi-head attention mechanism, enabling the model to focus on different parts of the input for a more nuanced representation. dim_{MLP} sets the dimensionality of the hidden layer in the multi-layer perceptron, to model more complex interactions while managing model size and computation. α_{common} indicates the dropout rate for the model, a regularization technique to prevent overfitting by randomly setting a fraction of the input units to zero during training. Similarly, α_{emb} defines the dropout rate applied to the embedding layer, also aiding in overfitting prevention.

Bayesian optimization differs from grid search or random search as it considers all historical evaluations [6]. This approach can be mathematically defined

as presented in Equation 1. The objective function is defined in the domain of X ; $f : X \rightarrow \mathbb{R}$.

$$\hat{x} \in \arg \max_{x \in X} f(x). \tag{1}$$

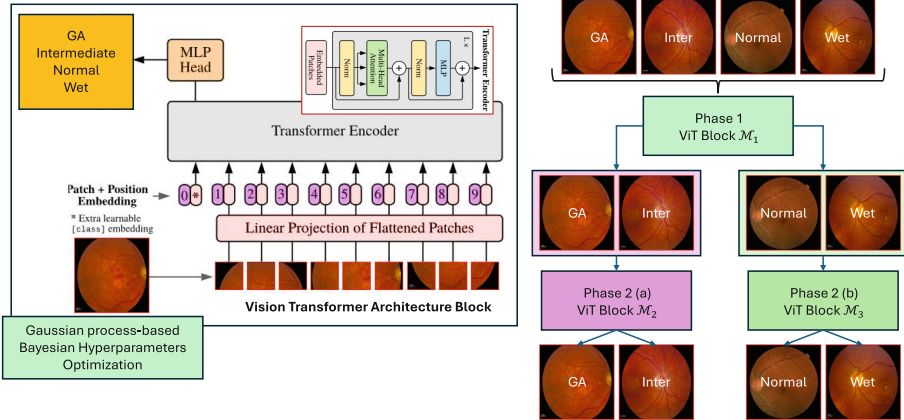


Fig. 3. Visualization of the utilized ViT architectures and the two cascaded phases using the three models (i.e., \mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3).

4.3 Architecture Explainability and Interpretability

In the context of AMD, the interpretability and explainability of machine learning models are crucial for ensuring reliable and understandable predictions. Several techniques can be employed to achieve this, including SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and GradCAM (Gradient-weighted Class Activation Mapping).

SHAP values provide a unified measure of feature importance by attributing the prediction of a model to its features based on cooperative game theory. The SHAP value ϕ_i for a feature i is calculated as in Equation 2 where N is the set of all features, S is a subset of N not containing feature i , and $f(S)$ is the prediction of the model with features in subset S .

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{2}$$

LIME explains individual predictions by locally approximating the model with an interpretable one. The key idea is to perturb the input data and observe the changes in the predictions. The explanation model g is trained to minimize the following objective as presented in Equation 3 where \mathcal{L} is the loss function

(e.g., mean squared error), f is the original model, π_x is the locality measure around instance x , and $\Omega(g)$ is the complexity of the explanation model g .

$$\xi(x) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (3)$$

GradCAM is used to generate visual explanations for CNN-based models by highlighting the regions of the input image that are important for the prediction. The importance of each pixel is determined by the gradients of the target class score to the feature maps of a convolutional layer. The GradCAM heatmap $L_{\text{Grad-CAM}}^c$ for class c is computed as in Equation 4 where A^k is the activation map of the k -th feature, and α_k^c is the weight for the k -th feature map, calculated as in Equation 5 where y^c is the score for class c , and Z is the number of pixels in the feature map.

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (4)$$

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (5)$$

In the diagnosis and treatment of AMD, these interpretability methods can help clinicians understand how machine learning models make decisions based on retinal images. For instance: (1) SHAP values can identify which features (e.g., drusen area, retinal thickness) are most influential in predicting AMD progression, (2) LIME can provide local explanations for individual patient predictions, helping clinicians understand specific model outputs, and (3) GradCAM can visualize the regions in retinal images that the model focuses on, assisting in the validation of automated diagnoses. *These methods enhance trust and transparency in AI-assisted medical diagnostics, contributing to more reliable and explainable healthcare solutions.*

5 Experiments and Discussion

Hardware and Software Setup: For this study, experiments were conducted on a local machine with an NVIDIA Quadro M4000 GPU (8GB) and 128GB CPU memory. Model development, training, validation, and testing were done using Python 3.9 [24] and PyTorch [16]. Performance metrics, including accuracy, sensitivity, and specificity, were calculated with SciKit-Learn [18]. Visualization of results, such as performance metrics, confusion matrices, feature extraction outcomes, and activation maps, was performed using Matplotlib [11] and SeaBorn [25]. Hyperparameter tuning was managed with Optuna, and eXplainable AI was achieved using SHAP, LIME, and GradCAM.

We evaluated batch sizes of 16, 32, and 64 to optimize computational efficiency and training stability. Images were resized to 224x224 pixels with a patch size of 16 to capture spatial information. Overfitting was prevented using early stopping and dynamically adjusted learning rates with the Adam optimizer [29],

based on Optuna’s recommendations. Through 200 trials, we identified optimal hyperparameters that maximized accuracy and precision.

Hyperparameter Tuning: The hyperparameter tuning experiments for ViT were executed for each batch size, with each execution involving 200 iterations. The chosen tuning method was the Bayesian Optimization algorithm. As mentioned in the methodology, we aimed to tune the dimensionality (dim), depth of transformer encoder layers (D), number of attention heads (H), MLP dimension (dim_{MLP}), dropout rates (α_{common} and α_{emb}), and learning rate (γ). The ranges of them are shown in Table 1.

Table 1. The utilized search space for the different hyperparameters in the current study.

Hyperparameter	Range	Step
Dimensionality (dim)	32 to 128	16
Depth of transformer encoder layers (D)	2 to 8	1
Number of attention heads (H)	2 to 8	1
MLP dimension (dim_{MLP})	32 to 128	16
Dropout rates (α_{common} and α_{emb})	0.1 to 0.5	0.05
Learning rate (γ)	10^{-5} to 10^{-3}	log
Batch size (BS)	16, 32 and 64	-

The hyperparameter tuning for ViT models without data augmentation reveals distinct configurations that lead to optimal model performance across various batch sizes. For \mathcal{M}_1 with a batch size of 16, the best trial was achieved with dim of 128, D of 3, H at 2, dim_{MLP} of 112, α_{common} of 0.2, α_{emb} of 0.1, and γ of approximately 0.000172. Increasing the batch size to 32, the model’s optimal hyperparameters shifted to a dim of 64, D of 6, H at 7, dim_{MLP} of 80, higher α_{common} of 0.4, α_{emb} of 0.5, and γ of roughly 0.000737. At a batch size of 64, the best-performing trial for \mathcal{M}_1 presented dim of 112, D of 8, H at 4, dim_{MLP} of 96, α_{common} of 0.3, α_{emb} of 0.15, and γ of about 0.000274.

For \mathcal{M}_2 with a batch size of 16, the trials that stood out featured dimensions ranging from 32 to 112, D s from 3 to 8, H from 4 to 7, and dim_{MLP} s from 96 to 128. The α_{common} s varied from 0.2 to 0.5, with α_{emb} s between 0.15 to 0.25, and γ s spanning from approximately 9.3E-05 to 0.000423. \mathcal{M}_2 with a batch size of 32 reached its peak with a dimension of 64, D of 6, H of 4, an dim_{MLP} of 96, α_{common} of 0.45, α_{emb} of 0.2, and γ of about 0.000630. Lastly, for \mathcal{M}_3 with a batch size of 16, the best trial showcased a dim of 48, D of 7, H at 3, an dim_{MLP} of 80, α_{common} of 0.45, α_{emb} of 0.15, and γ of approximately 0.000726.

5.1 Results and Discussion

We present final testing metrics for three models across batch sizes: 16, 32, and 64 without data augmentation. For batch size 16, \mathcal{M}_1 achieved 86.36% accu-

racy and perfect sensitivity. \mathcal{M}_2 had 95.45% accuracy and 94.74% sensitivity. \mathcal{M}_3 had 84.09% accuracy and 96.30% sensitivity. Increasing the batch size to 32 saw \mathcal{M}_1 's accuracy drop to 48.86%, though sensitivity remained perfect. \mathcal{M}_2 achieved 84.09% accuracy, and \mathcal{M}_3 achieved 77.27% accuracy with 79.17% sensitivity. At batch size 64, \mathcal{M}_1 had 85.23% accuracy and 95.12% sensitivity, \mathcal{M}_2 had 75.00% accuracy and 95.65% sensitivity, and \mathcal{M}_3 had 79.55% accuracy and 72.73% sensitivity. *Conclusion: A batch size of 16 is optimal, offering the highest accuracy, sensitivity, and lowest test loss, making it ideal for model development.*

Hyperparameter tuning for ViT models with data augmentation showed batch size 16 yielding the best performance, with accuracies from 0.8895 to 0.9535. Models, especially \mathcal{M}_1 and \mathcal{M}_3 , adapted well to various hyperparameters, including dimensions from 80 to 112, dropout rates of 0.1-0.2, and different depths and attention head structures. This highlights batch size 16's effectiveness in training dynamics and optimizing precision and sensitivity.

Batch size 32 balanced computational efficiency and model performance, achieving accuracies from 0.8837 to 0.9535 across various experiments. This size supported efficient gradient computations and enhanced convergence rates. Effective configurations included dimensions of 32 to 112, dropout rates of 0.2-0.45, and tailored depth and attention head structures. Notably, \mathcal{M}_2 performed well, underscoring batch size 32's role in optimizing training dynamics and model efficacy.

Batch size 64 offered computational benefits but showed variable performance, with accuracies from 0.6628 to 0.907. Models needed careful hyperparameter adjustments, such as dropout rates of 0.25-0.45 and dimensions from 80 to 128, to avoid overfitting or underfitting. Effective outcomes were seen in \mathcal{M}_2 and \mathcal{M}_3 , suggesting batch size 64's potential in specific contexts with tailored adjustments, but further refinement is needed for consistently high performance.

Evaluating three models across batch sizes revealed \mathcal{M}_1 excelled at batch size 16 with 95.45% accuracy, perfect sensitivity, and low test loss. \mathcal{M}_2 and \mathcal{M}_3 showed higher test losses and slightly lower accuracies. At batch size 32, \mathcal{M}_2 excelled with 95.45% accuracy and perfect sensitivity. \mathcal{M}_3 improved with fewer false negatives. At batch size 64, \mathcal{M}_3 achieved the highest accuracy (97.73%) and perfect sensitivity. These results suggest \mathcal{M}_1 performs well across all sizes, while \mathcal{M}_2 and \mathcal{M}_3 benefit from larger sizes.

In a cascade method, starting with models with high sensitivity is crucial to avoid missing positives early on. \mathcal{M}_1 with batch size 16 has perfect sensitivity and high accuracy (95.45%), making it suitable for the first stage. \mathcal{M}_2 with batch size 32 also has perfect sensitivity and accuracy, capturing all positives for further analysis. \mathcal{M}_3 with batch size 64 has the highest accuracy (97.73%) and perfect sensitivity, ideal for final stages to confirm true positives with minimal false positives. Testing all three sizes will help determine the best configuration.

We evaluated a cascading ensemble of \mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3 on a test dataset. The ensemble achieved 87.50% overall accuracy, with 93.48% recall for Normal or Wet conditions, showing proficiency in correctly classifying these instances. Sensitivity for GA or Intermediate conditions was 80.95%, indicating occasional

classification challenges. The confusion matrix showed the ensemble correctly identified many instances but also had some misclassifications.

Based on cascade method results with batch sizes 16, 32, and 64 with data augmentation, accuracy marginally improved with larger sizes (0.9205 for 16, 0.9318 for 32 and 64). Sensitivity varied between conditions and sizes. Batch size 32 had the highest sensitivity for Normal conditions (97.44%) but lower for Wet conditions (89.80%). Batch size 64 balanced high sensitivity for Wet conditions (94.44%) and respectable sensitivity for Normal conditions (91.18%). Batch size 16 had the lowest sensitivity. While batch size 32 excelled in Normal sensitivity, batch size 64 offered balanced performance, making it the best choice for overall sensitivity balance.

Table 2 summarizes the results for different configurations. Figure 4 shows hyperparameter optimization using Bayesian methods, which effectively utilized past results to identify optimal parameter combinations, leading to improved outcomes.

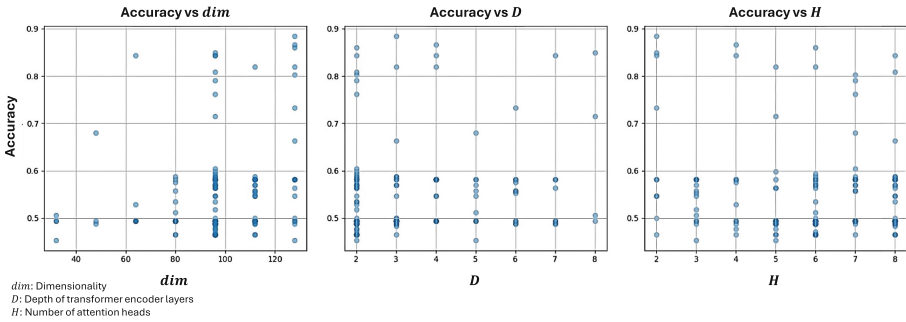


Fig. 4. Visual example of the tuning approach using Bayesian optimization for \mathcal{M}_1 with batch size of 16.

5.2 Explainability and Interpretability

As presented in Figure 5, we generated SHAP heat maps for random samples. These heat maps illustrate the importance of each pixel in the model’s prediction process. In our heat maps, blue pixels indicate features that decrease the model’s prediction (i.e., negative impact), while red pixels indicate features that increase the model’s prediction (i.e., positive impact). White pixels denote features that have no impact on the model’s prediction. *By overlaying these heat maps onto images, we can visualize which parts of the image are most influential for the model’s prediction. For instance, observing red areas near the optic disc might suggest their significance in the diagnostic process.*

Using LIME, we can utilize a deeper understanding of how our model interprets these regions. LIME generates local explanations by highlighting which regions within the fundus image contribute most significantly to the model’s

predictions. *The highlighted yellow outlines emphasize specific regions of interest, potentially crucial for medical diagnosis or research purposes.*

Table 2. Summarization of the obtained results for the different configurations. **Bold** results are for the cascaded suggested approach. **Blue** results are the best reported metrics.

Model	Loss	ACC (%)	REC (%)	SPEC (%)	Hyperparameters
\mathcal{M}_1	0.3081	86.36	100	87.72	$dim: 128, D: 3, H: 2, dim_{MLP}: 112, \alpha_{common}: 0.2, \alpha_{emb}: 0.1, \gamma: 0.00017, BS: 16$
\mathcal{M}_2	0.1609	95.45	94.74	94.74	$dim: 32, D: 6, H: 4, dim_{MLP}: 96, \alpha_{common}: 0.5, \alpha_{emb}: 0.15, \gamma: 0.00034, BS: 16$
\mathcal{M}_3	0.3089	85.23	95.12	85.77	$dim: 48, D: 7, H: 3, dim_{MLP}: 80, \alpha_{common}: 0.45, \alpha_{emb}: 0.15, \gamma: 0.00072, BS: 16$
\mathcal{M}_1	0.7705	48.86	100	100	$dim: 64, D: 6, H: 7, dim_{MLP}: 80, \alpha_{common}: 0.4, \alpha_{emb}: 0.5, \gamma: 0.00073, BS: 32$
\mathcal{M}_2	0.3944	84.09	100	85.11	$dim: 64, D: 6, H: 4, dim_{MLP}: 96, \alpha_{common}: 0.45, \alpha_{emb}: 0.2, \gamma: 0.00062, BS: 32$
\mathcal{M}_3	0.5292	77.27	79.17	79.17	$dim: 128, D: 8, H: 4, dim_{MLP}: 48, \alpha_{common}: 0.3, \alpha_{emb}: 0.2, \gamma: 0.00060, BS: 32$
\mathcal{M}_1	0.4771	85.23	95.12	85.77	$dim: 128, D: 6, H: 8, dim_{MLP}: 112, \alpha_{common}: 0.15, \alpha_{emb}: 0.3, \gamma: 0.00016, BS: 64$
\mathcal{M}_2	0.4834	75.00	95.65	79.90	$dim: 64, D: 2, H: 5, dim_{MLP}: 48, \alpha_{common}: 0.4, \alpha_{emb}: 0.25, \gamma: 0.00092, \alpha_{emb}: 0.4, \gamma: 0.0008, BS: 64$
\mathcal{M}_3	0.4001	79.55	72.73	78.05	$dim: 128, D: 2, H: 8, dim_{MLP}: 112, \alpha_{common}: 0.2, \alpha_{emb}: 0.3, \gamma: 0.00067, BS: 64$
\mathcal{M}_1	0.0786	95.45	100	91.00	$dim: 112, D: 2, H: 7, dim_{MLP}: 32, \alpha_{common}: 0.15, \alpha_{emb}: 0.15, \gamma: 0.00012, BS: 16, With DA$
\mathcal{M}_2	0.3002	93.18	95.83	90.00	$dim: 80, D: 3, H: 7, dim_{MLP}: 128, \alpha_{common}: 0.25, \alpha_{emb}: 0.4, \gamma: 0.00065, BS: 16, With DA$
\mathcal{M}_3	0.3255	93.18	84.21	88.00	$dim: 80, D: 5, H: 2, dim_{MLP}: 48, \alpha_{common}: 0.3, \alpha_{emb}: 0.1, \gamma: 0.00012, BS: 16, With DA$
Cascaded	-	92.05	93.33	90.70	BS: 16, With DA
\mathcal{M}_1	0.1632	94.32	92.68	91.00	$dim: 32, D: 2, H: 6, dim_{MLP}: 32, \alpha_{common}: 0.2, \alpha_{emb}: 0.1, \gamma: 0.00095, BS: 32, With DA$
\mathcal{M}_2	0.1681	95.45	100	91.00	$dim: 112, D: 6, H: 5, dim_{MLP}: 64, \alpha_{common}: 0.45, \alpha_{emb}: 0.35, \gamma: 0.00099, BS: 32, With DA$
\mathcal{M}_3	0.3131	93.18	94.44	92.00	$dim: 80, D: 4, H: 5, dim_{MLP}: 112, \alpha_{common}: 0.1, \alpha_{emb}: 0.15, \gamma: 0.00031, BS: 32, With DA$
Cascaded	-	93.18	89.80	97.44	BS: 32, With DA
\mathcal{M}_1	0.1996	93.18	91.11	91.00	$dim: 80, D: 3, H: 7, dim_{MLP}: 128, \alpha_{common}: 0.25, \alpha_{emb}: 0.4, \gamma: 0.00065, BS: 64, With DA$
\mathcal{M}_2	0.0842	95.45	94.12	96.00	$dim: 80, D: 2, H: 2, dim_{MLP}: 48, \alpha_{common}: 0.3, \alpha_{emb}: 0.25, \gamma: 0.00021, BS: 64, With DA$
\mathcal{M}_3	0.1167	97.73	100	94.00	$dim: 80, D: 7, H: 3, dim_{MLP}: 112, \alpha_{common}: 0.25, \alpha_{emb}: 0.15, \gamma: 0.0002, BS: 64, With DA$
Cascaded	-	93.18	94.44	91.18	BS: 64, With DA

ACC: Accuracy, REC: Recall, and SPEC: Specificity.

Grad-CAM generated heatmaps that highlight regions crucial for the ViT’s decision-making process. In our visualization, warmer colors such as red and yellow indicate areas where the model assigns higher importance. Conversely, cooler colors, particularly blue, signify regions where the model assigns less importance. *These blue areas suggest that certain parts of the fundus images are deemed less critical by the ViT in its classification or diagnostic process.*

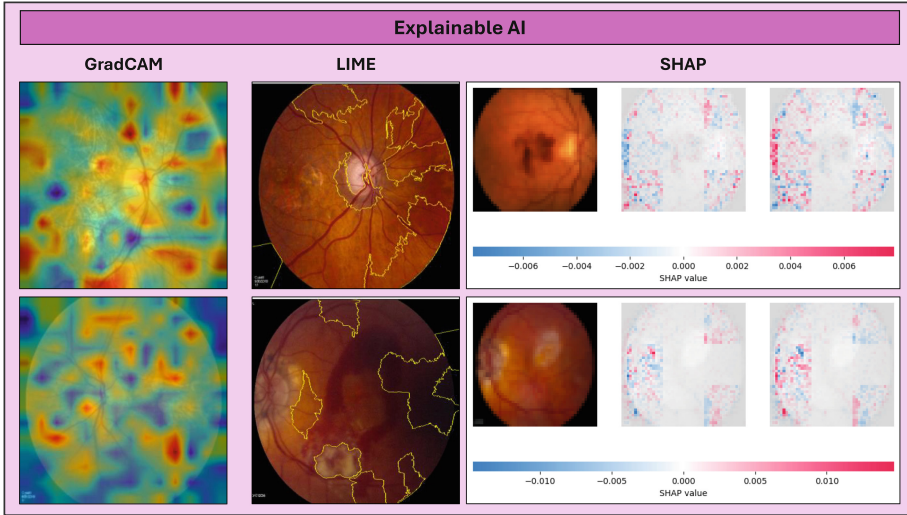


Fig. 5. Explainability and interpretability for AMD diagnosis. SHAP heat maps show pixel importance (blue: negative, red: positive, white: neutral). LIME highlights significant regions (yellow outlines) in fundus images. Grad-CAM heatmaps reveal key areas for ViT’s decisions (red/yellow: high importance, blue: low importance).

5.3 Ablation Studies

First Ablation Study: Removal of the Cascaded Approach: In the first ablation study, the cascaded approach was removed, and all four classes were simultaneously applied to the ViT. The results demonstrate performance metrics with a higher test loss of 0.5744 and a lower Accuracy of 80.68% compared to the cascaded method. *This indicates that applying all classes together was ineffective, despite having sensitivities for each class as follows: 1.0, 0.636, 0.913, and 0.65. Specificity values for each class are: 0.862, 0.955, 0.969, and 0.956.*

Second Ablation Study: Removal of the Tuning Process: In the second ablation study, the ViT model underwent a random experiment without tuning. The results indicate a significant decrease in performance compared to the tuned model, with a higher test loss of 0.7509 and a lower Accuracy of 69.32%. *This highlights the critical role of tuning in optimizing ViT models for specific tasks, as untuned models may struggle with accurate classification.*

5.4 Comparison with Related Studies

Our study evaluated a cascading ensemble of models \mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3 , achieving an overall accuracy of 93.18%. Sensitivity for class A was 94.44% and for class B was 91.18%, indicating robust performance in correctly identifying both classes. Compared to Gholami et al. [9], who achieved accuracy rates of 98.18% and 99.11% with centralized ViT models, our accuracy is slightly lower. Similarly,

Xu et al. [26] achieved AUC values of 98.76% and 96.47% with DeepDrAMD, while our sensitivity metrics are comparable, suggesting strong but slightly lower performance. Yao et al. [27] reported an accuracy of 98.66% and an F1 score of 98.96% for macular edema, indicating that our model’s accuracy of 93.18% shows potential but requires improvement to reach such a high performance.

Moreover, Kihara et al. [13] reported sensitivities of 82% and specificities of 90% for neMNV detection, which are lower than our sensitivities, though focused on different conditions. Akcca et al. [5] achieved higher accuracies with different ViT models, notably 99.17% with Mobile-ViT, highlighting areas where our ensemble approach could benefit from different transformer architectures. Finally, Jiang et al. [12] achieved an impressive 99.69% accuracy with a pruned ViT, far surpassing our 93.18%, underscoring the exceptional performance of optimized ViT models. *Overall, while our cascading ensemble shows promising results, there is room for improvement, particularly by exploring different transformer architectures and optimizing model configurations to match state-of-the-art performance.*

6 Limitations

The study faced limitations that impacted outcomes and model generalizability. The dataset used consisted of OCT fundus images from a small collection. Data scarcity was particularly problematic for rare conditions like age-related macular degeneration, further hindering model performance. To address this, data augmentation techniques were used, aiming to enhance performance but risking overfitting and increased computing demands. Balancing augmentation and computing resources was crucial to optimize model performance. A larger, diverse dataset could have better represented retinal diseases, improving model robustness and generalizability.

7 Conclusions and Future Directions

AMD causes vision loss in older adults and requires timely diagnosis. With an aging population, advanced diagnostic tools are essential. AI, through automated image analysis, can improve AMD detection, enhancing early intervention and personalized treatment. Our cascading ensemble of models (\mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3) achieved 87.50% overall accuracy, with high sensitivity (93.48%) for Normal or Wet conditions and lower sensitivity (80.95%) for GA or Intermediate conditions. Data augmentation and varying batch sizes showed impacts on performance: batch size 16 achieved 92.05% accuracy without augmentation, while batches 32 and 64 marginally improved to 93.18%. Batch size 32 had the highest sensitivity for Normal conditions (97.44%) but lower for Wet conditions (89.80%). Batch size 64 offered balanced sensitivity, with 94.44% for Wet and 91.18% for Normal conditions. Thus, batch size 64 was optimal for balanced sensitivity across conditions. Future research should focus on improving sensitivity for GA and Intermediate conditions, possibly by developing specialized models or

incorporating additional features. Advanced data augmentation techniques, such as using GANs to generate high-quality synthetic data, should be explored to mitigate overfitting and improve performance, especially for rare conditions.

References

1. CATT — Center for Preventive Ophthalmology and Biostatistics (CPOB) — Perelman School of Medicine at the University of Pennsylvania — [med.upenn.edu](https://www.med.upenn.edu/cpob/catt.html). <https://www.med.upenn.edu/cpob/catt.html>, [Accessed 18-06-2024]
2. Abd El-Khalek, A.A., Balaha, H.M., Mahmoud, A., Alghamdi, N.S., Ghazal, M., Khalil, A.T., Abo-Elvoud, M.E.A., El-Baz, A.: A novel machine learning-based classification framework for age-related macular degeneration (amd) diagnosis from fundus images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2024)
3. Abd El-Khalek, A.A., Balaha, H.M., Sewelam, A., Ghazal, M., Khalil, A.T., Abo-Elvoud, M.E.A., El-Baz, A.: A comprehensive review of ai diagnosis strategies for age-related macular degeneration (amd). *Bioengineering* **11**(7) (2024)
4. Abdin, A.D., Devenijn, M., Fulga, R., Langenbacher, A., Seitz, B., Kaymak, H.: Prevalence of geographic atrophy in advanced age-related macular degeneration (amd) in daily practice. *J. Clin. Med.* **12**(14), 4862 (2023)
5. Akça, S., Garip, Z., Ekinçi, E., Atban, F.: Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal oct images using vision transformers: a comparative study. *Lasers Med. Sci.* **39**(1), 140 (2024)
6. Bai, T., Li, Y., Shen, Y., Zhang, X., Zhang, W., Cui, B.: Transfer learning for bayesian optimization: A survey. *arXiv preprint [arXiv:2302.05927](https://arxiv.org/abs/2302.05927)* (2023)
7. Blasiak, J., Pawłowska, E., Ciupińska, J., Derwich, M., Szczepanska, J., Kaarniranta, K.: A new generation of gene therapies as the future of wet amd treatment. *Int. J. Mol. Sci.* **25**(4), 2386 (2024)
8. Elgafi, M., Sharafeldeen, A., Elnakib, A., Elgarayhi, A., Alghamdi, N.S., Sallah, M., El-Baz, A.: Detection of diabetic retinopathy using extracted 3d features from oct images. *Sensors* **22**(20), 7833 (2022)
9. Gholami, S., Lim, J.I., Leng, T., Ong, S.S.Y., Thompson, A.C., Alam, M.N.: Federated learning for diagnosis of age-related macular degeneration. *Frontiers in Medicine* **10** (2023)
10. Haggag, S., Elnakib, A., Sharafeldeen, A., Elsharkawy, M., Khalifa, F., Farag, R.K., Mohamed, M.A., Sandhu, H.S., Mansoor, W., Sewelam, A., El-Baz, A.: A computer-aided diagnostic system for diabetic retinopathy based on local and global extracted features. *Appl. Sci.* **12**(16), 8326 (2022). <https://doi.org/10.3390/app12168326>
11. Hunter, J.D.: Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* **9**(3), 90–95 (2007). <https://doi.org/10.1109/MCSE.2007.55>
12. Jiang, Z., Wang, L., Wu, Q., Shao, Y., Shen, M., Jiang, W., Dai, C.: Computer-aided diagnosis of retinopathy based on vision transformer. *Journal of Innovative Optical Health Sciences* **15**(02), 2250009 (2022)
13. Kihara, Y., Shen, M., Shi, Y., Jiang, X., Wang, L., Laiginhas, R., Lyu, C., Yang, J., Liu, J., Morin, R., et al.: Detection of nonexudative macular neovascularization on structural oct images using vision transformers. *Ophthalmology Science* **2**(4), 100197 (2022)

14. Lee, G., et al.: Parallel vs. sequential cascading mep coordination strategies: A pharmaceutical building case study. *Automation in Construction* **43**, 170–179 (2014)
15. Liu, Y., Jing, W., Xu, L.: Parallelizing backpropagation neural network using mapreduce and cascading model. *Comput. Intell. Neurosci.* **2016**(1), 2842780 (2016)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems* 32, pp. 8024–8035. Curran Associates, Inc. (2019)
17. Paul, S., Chen, P.Y.: Vision transformers are robust learners. In: *Proceedings of the AAAI conference on Artificial Intelligence*. vol. 36, pp. 2071–2081 (2022)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
19. Rane, N.: Transformers for medical image analysis: Applications, challenges, and future scope. *Challenges, and Future Scope* (November 2, 2023) (2023)
20. Sabry, M., Balaha, H.M., Ali, K.M., Soliman, T.H.A., Gondim, D., Ghazal, M., Tahtouh, T., El-Baz, A.: A vision transformer approach for breast cancer classification in histopathology. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. pp. 1–4. IEEE (2024)
21. Schultz, N.M., Bhardwaj, S., Barclay, C., Gaspar, L., Schwartz, J.: Global burden of dry age-related macular degeneration: a targeted literature review. *Clin. Ther.* **43**(10), 1792–1818 (2021)
22. Sharafeldeen, A., Elgafi, M., Elnakib, A., Mahmoud, A., Elgarayhi, A., Alghamdi, N.S., Sallah, M., El-Baz, A.: Diabetic retinopathy detection using 3d oct features. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE (Apr 2023<https://doi.org/10.1109/isbi53787.2023.10230785>)
23. Sun, W., Zhao, Y., Liao, L., Wang, X., Wei, Q., Chao, G., Zhou, J.: Effects of acupuncture on age-related macular degeneration: A systematic review and meta-analysis of randomized controlled trials. *PLoS ONE* **18**(3), e0283375 (2023)
24. Van Rossum, G., Drake, F.L.: *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA (2009)
25. Waskom, M.L.: seaborn: statistical data visualization. *Journal of Open Source Software* **6**(60), 3021 (2021<https://doi.org/10.21105/joss.03021>, <https://doi.org/10.21105/joss.03021>)
26. Xu, K., Huang, S., Yang, Z., Zhang, Y., Fang, Y., Zheng, G., Lin, B., Zhou, M., Sun, J.: Automatic detection and differential diagnosis of age-related macular degeneration from color fundus photographs using deep learning with hierarchical vision transformer. *Comput. Biol. Med.* **167**, 107616 (2023)
27. Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., Wang, Z.: Funswin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. *Front. Physiol.* **13**, 961386 (2022)
28. de Zarzà, I., de Curtò, J., Hernández-Orallo, E., Calafate, C.T.: Cascading and ensemble techniques in deep learning. *Electronics* **12**(15), 3354 (2023)
29. Zhang, Z.: Improved adam optimizer for deep neural networks. In: *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. pp. 1–2. Ieee (2018)



DCRUNet++: A Depthwise Convolutional Residual UNet++ Model for Brain Tumor Segmentation

Yash Sonawane¹, Maheshkumar H. Kolekar², Agnesh Chandra Yadav^{2(✉)},
Gargi Kadam¹, Sanika Tiwarekar¹, and Dhananjay R. Kalbande¹

¹ Bharatiya Vidya Bhavans Sardar Patel Institute of Technology, Mumbai, India

² Indian Institute of Technology Patna, Bihar, India

agnesh_2221ee22@iitp.ac.in

Abstract. The rapid advancement of emerging technologies is reshaping brain tumor diagnosis and treatment planning, with a focus on precise segmentation techniques for early intervention. Manual segmentation methods face challenges due to inherent noise and intensity variations in medical imaging data. To mitigate these challenges, we propose DCRUNet++ for brain tumor segmentation. The proposed model integrates Depthwise convolutional residual module blocks to enhance information flow and gradient propagation across network layers, thereby improving feature representation. The DCRUNet++ architecture incorporates nested up-convolution operations, facilitating the propagation of semantic information from lower to higher levels of abstraction. To further optimize model training, we introduce a custom loss function that assigns higher weights to feature maps 4 and 8, prioritizing significant representations during the optimization process. Deep supervision, utilizing 8 intermediate feature maps, ensures robust training and facilitates convergence by emphasizing critical representations. Extensive experimentation with FLAIR MRI images validates the efficacy of the proposed DCRUNet++ model. Achieving a Dice coefficient of 0.9467 and a mean Intersection over Union of 0.9155, our model outperforms previous methodologies, underscoring its effectiveness in brain tumor segmentation and treatment planning.

Keywords: Brain tumor segmentation · DCRUNet++ · FLAIR MRI · UNet++

1 Introduction

Morphological changes in brain tissue segmentation are essential for evaluating the progression of neurological disorders. Brain and CNS tumors have the highest mortality rates among these disorders. Accurate brain tumor segmentation using neuroimaging is crucial for improving diagnosis, treatment, monitoring, and research. Brain tumors' variability in location, shape, and size makes segmentation challenging. Medical imaging modalities like CT, MRI, and PET are

used for identification, with MRI preferred for its superior tissue resolution. The quality of brain cancer treatment largely depends on the physician's expertise due to the complex nature of brain tumors [17]. MRI is critical in glioma diagnosis, utilizing several protocols, including T1c and commonly used sequences like T1-weighted, T2-weighted, and FLAIR [22]. These modalities offer unique tissue contrast for comprehensive tumor assessment and segmentation, essential for delineating tumor boundaries, assessing heterogeneity, and planning targeted therapies. Advanced machine learning (ML) and deep learning (DL) algorithms are increasingly used in CAD systems to enhance segmentation accuracy and diagnostic precision [9].

Recently, ML models have been widely used for object prediction and classification in healthcare, particularly for forecasting pandemics and disorders [1, 6]. Research focuses on applying ML to characterize brain tumor images. Traditional ML approaches use manually selected features, while state-of-the-art DL models use multiple layers and functions to automatically extract features from raw data, enhancing classification, segmentation, and image analysis. DL models, especially CNNs, have shown superior performance in detecting and segmenting brain tumors. Both ML and DL techniques are increasingly adopted to improve diagnostic accuracy and efficiency in medical imaging [14, 27, 32].

In this paper, we employed a novel approach utilizing FLAIR MRI images for segmentation by incorporating a Multi-Scale Residual Module into a UNet++ architecture. FLAIR imaging is gaining significance in the segmentation of malignant tumors due to the trend of resecting FLAIR-positive areas. FLAIR effectively delineates the tumor boundaries, aiding in accurate surgical planning and identifying positive tumor regions. The segmentation model is evaluated on an independent dataset, and its performance is compared against various existing segmentation models using different statistical parameters.

The main contributions of the methodologies presented and extensively discussed in this study are outlined as follows:

1. We proposed a Depthwise Convolutional Residual Module (DCRM) , which enhances the propagation of gradients and optimizes the flow of information across the network layers.
2. Designed a DCRUNet++ model incorporating DCRM blocks, drawing inspiration from the UNet# architecture, featuring nested up-convolution operations. This enhancement facilitates enhanced semantic information transmission from lower to higher levels.
3. Developed a custom loss function that assigns higher weights to maps 4 and 8, ensuring the model focuses more on significant feature maps during training, leading to improved accuracy and performance.
4. Integrated Deep Supervision (DS) through the use of 8 intermediate feature maps. This approach, along with the custom loss function, ensures robust training and better convergence by emphasizing crucial intermediate representations.
5. Conducted extensive experiments and validations using FLAIR MRI images, demonstrating the effectiveness of the proposed model in accurately segmenting brain tumors.

The methodology of the proposed approach is elucidated in Section 2, followed by experimental analysis in Section 3, which illustrates the results obtained from the proposed model. Finally, Section 4 encapsulates the research findings and delineates avenues for future work.

1.1 Related Work

This chapter provides an in-depth examination of the methodologies employed for diagnosing brain tumors using advanced technological approaches. Traditional context-based machine learning methods often fall short in tasks such as semantic segmentation, where deep learning algorithms have demonstrated superior performance[8]. Among medical imaging techniques, image segmentation is extensively utilized for the automated identification and delineation of tissues and pathologies.

In image classification tasks, benchmark models such as AlexNet [16], VGGNet [26], and GoogleNet [29] have achieved notable success. However, deeper networks face degradation issues, which can be mitigated by using skip connections [15]. Skip connections allow models to reach greater depth by transferring information from initial to deeper layers without adding parameters, as shown in ResNets [11]. DenseNets [12] enhance this by concatenating feature maps from earlier layers with those from deeper layers, promoting feature reusability while using fewer parameters. This results in deeper representations and faster convergence. Advancements in skip connections have significantly improved performance in UNet-derived models like UNet2+ [33] and UNet3+ [13]. UNet, a successful neural network for healthcare image segmentation, utilizes skip connections to concatenate the encoder's feature map with the decoder's upsampled feature map, enhancing segmentation accuracy. Despite their success, UNet2+ and similar models may lack comprehensive information from full-scale investigations, affecting their precision in learning organ locations and boundaries. Using ResNets and DenseNets as backbones for the UNet encoder further enhances organ and lesion segmentation accuracy.

DS mitigates the vanishing gradient problem and accelerates convergence in deep neural networks [19]. Inception-v4 [28] incorporates auxiliary supervision classifiers in intermediate layers, enhancing training and reducing overfitting. Wang et al. [31] discuss the integration of DS to boost performance, widely applied in UNet-like segmentation networks. Li et al. [20] expanded UNet to eight layers, introducing an auxiliary pathway for mid-layer semantic supervision, improving left ventricle segmentation. Farheen et al. [7] developed MultiResUNet, incorporating DS in the decoder branch, reducing false negatives in tumor detection. Liu et al. [21] applied 3D DS in DSSE-V-Net, enhancing feature discrimination for brain tumor segmentation and quickening convergence. UNet3+ [13] leverages DS for hierarchical learning, improving organ and tissue segmentation at various scales. UNet2+ [33] uses DS for model pruning, speeding up inference with minimal performance loss.

Buda et al. [3] used deep learning-based segmentation to link genomic subtypes of low-grade gliomas (LGG) with tumor imaging features, achieving a high

Dice coefficient of 82%. Naser et al. [24] demonstrated UNet models for brain tumor segmentation, highlighting the need for large annotated datasets and significant computational resources. Walsh et al. [30] introduced a lightweight UNet for MRI tumor segmentation on the BITE dataset, achieving an IoU of 89%, suitable for real-time use but potentially less accurate for complex tumor structures.

Cinar et al. [5] proposed using DenseNet 121 as a backbone for a UNet architecture targeting the segmentation of FLAIR MRI images. Isaza et al. [2] investigated transfer learning for brain tumor segmentation, comparing various data augmentation methods on the ResNet50 network. This method achieved an impressive F1 detection score of 92.34% but may face challenges in adapting to diverse MRI imaging modalities or capturing fine-grained tumor boundaries due to inherent feature representation limitations of ResNet50. Ruba et al. [25] employed FCN for tumor localization and UNet for sub-region segmentation. While this approach offers flexibility in handling spatial dependencies, it may struggle with capturing contextual information across different regions of interest, potentially leading to segmentation errors. Kumar et al. [17] introduced residual models within a UNet architecture for brain tissue segmentation, showing improved performance using the FLAIR modality. However, the inclusion of residual models adds additional hyperparameters and complexity, necessitating careful tuning and optimization to achieve optimal results. Metlek et al. [23] proposed ResUNet+, leveraging convolution for regions of interest detected in different modalities. Despite its potential for multi-modal image segmentation, ResUNet+ faces increased computational complexity and training time, limiting its scalability to large-scale datasets.

2 Methodology

2.1 Overview

The proposed methodology introduces an innovative architecture, termed DCRUNet++, for brain tumor segmentation using FLAIR MRI images, as shown in Fig. 1. This approach leverages a *UNet#* inspired structure, enhanced with advanced features to improve segmentation accuracy and detail. The architecture incorporates additional nested up-convolution operations to provide more semantic and detailed information from lower to higher levels. A DCRM is proposed, as shown in Fig. 1(a), which includes internal skip connections integrated into the network and connected with dense skip connections to facilitate better information flow and feature maps. Additionally, DS is utilized for faster convergence and proper selection of intermediate feature maps, ensuring guided learning at multiple levels. A custom loss function prioritizes the 4th and 8th feature maps by assigning higher weights to these maps, with the sum of losses from all 8 maps used to adjust the model weights. This combination of architectural enhancements, DS, and a tailored loss function promotes balanced and effective training, resulting in a robust solution for accurate and detailed brain tumor segmentation using FLAIR MRI images.

2.2 Proposed Architecture

UNet++ serves as the backbone of the proposed DCRUNet++ model for brain tumor segmentation. The UNet++ architecture comprises a series of feature maps that represent progressively deeper network layers, capturing increasingly refined representations of the input data.

DCRUNet++ enhances this architecture by replacing standard convolutional layers with novel DCRMs across all encoder, decoder, and intermediate nodes. These DCRMs perform convolutions, batch normalization, and swish activations, integrated with skip connections to facilitate gradient flow and improve learning efficiency. Feature maps are extracted at different depths, with $DCRM^{0,0}$ to $DCRM^{0,4}$ representing the initial row of refined representations, and deeper layers like $DCRM^{1,0}$ to $DCRM^{1,3}$, $DCRM^{2,0}$ to $DCRM^{2,2}$, and $DCRM^{3,0}$ to $DCRM^{3,1}$ providing progressively abstract and high-level features, culminating in $DCRM^{4,0}$, the deepest feature map, illustrated in Fig. 1(b).

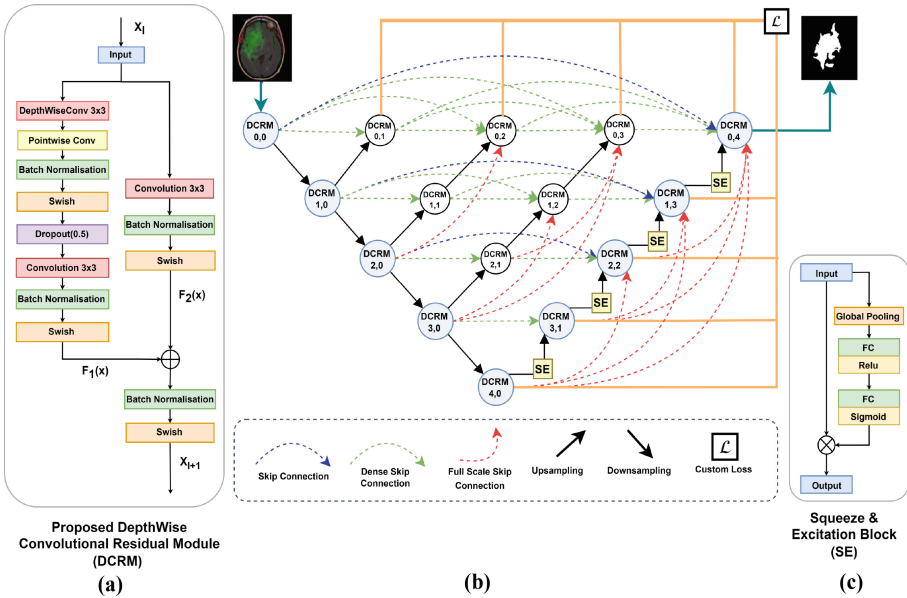


Fig. 1. (a) DepthWise Convolutional Residual Module Blocks (b) Overall Architecture of the Proposed DCRUNet++ Model (c) Squeeze-and-Excitation Block within DCRUNet++ Model

Dense skip connections aggregate features across all layers, while standard skip connections directly link corresponding encoder and decoder layers to retain high-resolution details. The model employs learnable upsampling using Conv2D transpose layers, which increase spatial dimensions during training. Squeeze-and-Excitation (SE) blocks, as shown in Fig. 1(c), depicted as yellow boxes,

recalibrate channel-wise feature responses, enhancing feature representation. The decoder’s final output, $DCRM^{0,4}$, is achieved by selecting the optimal channel from eight feature maps. This process integrates dense skip interconnections from $DCRM^{0,0}$ to $DCRM^{0,3}$, full-scale skip intraconnections from $DCRM^{4,0}$, $DCRM^{3,1}$, $DCRM^{2,2}$, and the Conv2D transpose of $DCRM^{1,3}$. The restructured skip connections enhance the similarity between encoder and decoder features, aiding the optimizer during training and boosting the model’s learning capability.

Full-scale skip connections bridge deep and shallow layers in the intermediate and decoder sub-networks, ensuring efficient gradient flow and feature reuse. DS is implemented by applying multiple loss functions (\mathcal{L}) at various intermediate layers, compelling the network to produce useful features at multiple scales and depths. The final output, a segmentation map of the brain tumor, is generated by integrating dense skip interconnections and full-scale skip intraconnections from various layers, resulting in comprehensive feature utilization. This architecture efficiently combines residual learning, dense connectivity, learnable upsampling, SE blocks, and DS to achieve high-performance brain tumor segmentation.

2.2.1 Depthwise Convolutional Residual Module The architectural design of the proposed residual module blocks, depicted in the Fig 1 (a), processes the input X_l through two parallel branches and subsequently merges their results.

In the left branch, the input initially undergoes a depthwise convolution with a 3x3 kernel, performing convolutions independently on each input channel. This layer is computationally efficient and excels at capturing spatial characteristics. Next, a pointwise convolution (1x1) merges features across a wide range of channels, significantly enhancing the depthwise convolution’s output by integrating information across the complete feature map, leading to a more comprehensive representation. We then introduce batch normalization as a regularization technique to prevent overfitting. We utilize the Swish activation function, denoted as:

$$\text{Swish}(x) = x \cdot \sigma(x) \quad (1)$$

This activation function combines linear and non-linear characteristics through the multiplication of the input variable x with its sigmoid activation, resulting in a continuous and non-monotonic change. Our experiments demonstrated that Swish enhances model performance compared to conventional activation functions such as ReLU. Furthermore, we incorporate a dropout layer with a 50% rate, which randomly deactivates half of the neurons during each training cycle. This layer helps alleviate overfitting by encouraging the network to learn more reliable features. The output then goes through an additional 3x3 convolution layer to further refine the feature representation, followed by another round of batch normalization and Swish activation.

In the right branch, the input undergoes a standard 3x3 convolution layer, primarily focusing on acquiring spatial features, followed by batch normalization. The outputs from both branches, $F_1(x)$ and $F_2(x)$, are then merged through

element-wise addition. This approach effectively mitigates the vanishing gradient issue. We introduce a batch normalization layer for the combined result to standardize the activations. Finally, a Swish activation function is applied to produce the final output, X_{l+1} .

2.2.2 Deep Supervision

DS is integrated into the proposed model architecture, allowing it to operate in two distinct modes: (1) accurate mode, where the outputs of the model’s branches are summed to compute the loss, and (2) fast mode, which selects the branch with the optimal result and prunes the model to enhance its speed. As shown in Fig. 2, the loss function is applied to the outputs of the 8 branches in the proposed DCRUNet++ architecture. Specifically, the implementation proceeds as follows: first, the dense skip interconnections and full-scale skip intraconnections generate the first layer feature maps $DCRM^{0:j}$ with $j \in [1, 2, 3]$ and $DCRM^{0,4}$ at multiple semantic levels.

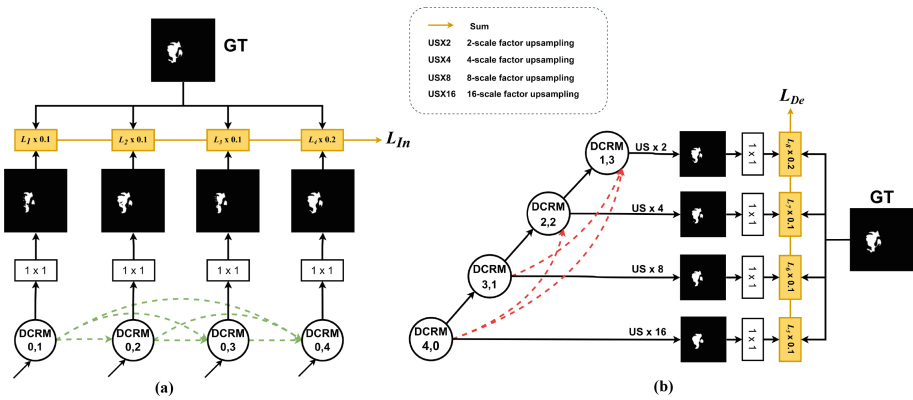


Fig. 2. The visual flowchart of DS in the proposed DCRUNet++ model demonstrates that DS serves two functions: (a) model pruning and (b) improvement of learning for hierarchical representations. Adding a loss function to each branch, followed by a 1×1 convolution operation, produces the loss \mathcal{L}^i , where $i \in [1, 4]$. \mathcal{L}_{In} is utilized to supervise model pruning, while \mathcal{L}_{De} is used to enhance feature learning in the model. It is noteworthy that these two DS forms, (a) and (b), employ distinct implementation approaches.

The feature maps are processed through a 1×1 convolution followed by a ReLU activation function, producing the output results. By comparing these results with the ground truth (GT), the loss \mathcal{L}_{In} is computed to supervise the model for pruning. Additionally, the decoder branches $DCRM^{1,3}$, $DCRM^{2,2}$, $DCRM^{3,1}$, and $DCRM^{4,0}$ are also processed by a 1×1 convolution and ReLU activation function, achieving their respective final results. Comparing these results with the appropriately down-scaled GT, the loss \mathcal{L}_{De} is calculated, facilitating DS for model training. This DS approach enhances the learning of hierarchical

representations from full-scale feature maps, ensuring robust feature extraction and effective gradient flow throughout the network.

2.3 Loss Function

In the proposed brain tumor segmentation methodology, we employ an integrated loss function framework that synergistically combines Binary Cross Entropy (BCE) loss and Dice loss. This approach aims to optimize both segmentation accuracy and spatial overlap, which are essential for the precise delineation of brain tumor regions in MRI scans.

The loss function $\mathcal{L}(M_i, T)$ for each intermediate feature map M_i incorporates both BCE loss and Dice loss to exploit their complementary strengths, thereby enhancing the performance of the deep learning model.

The combined loss $\mathcal{L}(M_i, T)$ is formulated as:

$$\mathcal{L}(M_i, T) = \text{BCE} + \alpha \cdot \text{Dice} \quad (2)$$

where α is a hyperparameter that adjusts the contribution of the Dice loss to the overall loss. In our implementation, α is set to 0.001, assigning a minor weight to the Dice loss while predominantly relying on the BCE loss for training stability. The feature maps M_1, M_2, M_3 , and M_4 are derived from the initial layers, denoted as $DCRM^{0,1}, DCRM^{0,2}, DCRM^{0,3}$, and $DCRM^{0,4}$, respectively. The subsequent feature maps M_5, M_6, M_7 , and M_8 are obtained from later layers, specifically $DCRM^{4,0}, DCRM^{3,1}, DCRM^{2,2}$, and $DCRM^{1,3}$, respectively as shown in Fig. 2.

Let $\{M_i\}_{i=1}^8$ represent the intermediate feature maps produced by the network, and $\mathcal{L}(M_i, T)$ denote the loss function computed between the intermediate feature map M_i and the ground truth target T . The weights assigned to the losses from these intermediate stages are as follows:

- The 4th and 8th feature maps (M_4 and M_8) are assigned a weight of 0.2 each.
- The remaining six feature maps (M_1, M_2, M_3, M_5, M_6 , and M_7) are assigned a weight of 0.1 each.

The final loss $\mathcal{L}_{\text{final}}$ is computed as the weighted sum of the individual losses from each of these intermediate feature maps. Mathematically, this is expressed as:

$$\mathcal{L}_{\text{final}} = \sum_{i=1}^8 w_i \cdot \mathcal{L}(M_i, T) \quad (3)$$

where the weights w_i are defined as:

$$w_i = \begin{cases} 0.2 & \text{if } i = 4 \text{ or } i = 8 \\ 0.1 & \text{otherwise} \end{cases} \quad (4)$$

Consequently, these losses receive greater emphasis in the overall loss computation. Such an approach facilitates iterative refinement of model predictions across various stages, fostering improved convergence and performance through effective utilization of intermediate supervision signals.

3 Experiment

3.1 Dataset

We leveraged the TCGA-LGG dataset from Kaggle, originally sourced from The Cancer Imaging Archive (TCIA). This comprehensive dataset includes 3929 images derived from 110 participants, classified into 1373 images labeled as class '1' (Tumor) and 2556 images labeled as class '0' (Normal). For training our proposed model, we designated 2750 images for training, 589 images for testing, and 590 images for validation. Each image in the dataset is a 2D representation with precisely defined FLAIR abnormality masks, each measuring 256 x 256 pixels. Representative samples of the dataset are displayed in Fig. 3.

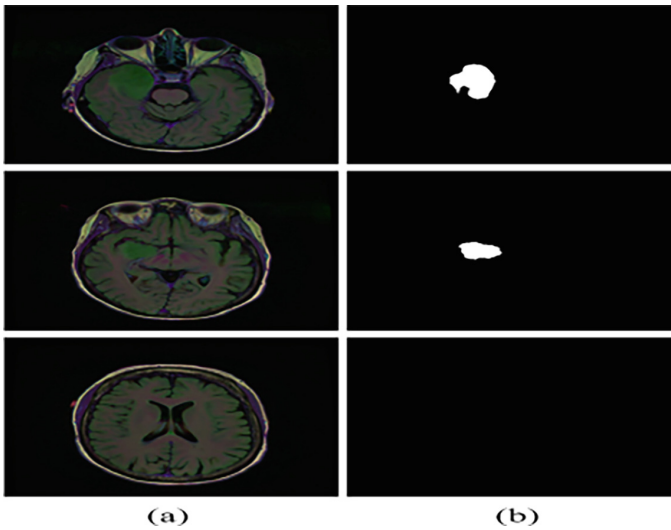


Fig. 3. Few samples of dataset. (a) Original Image (b) Original Mask

3.2 Implementation Details

We employ a variety of libraries and frameworks for image segmentation and processing, including segmentation-models-pytorch and OpenCV. To enhance the diversity of our dataset, we use Augmentor, while SciPy's morphology module is applied for morphological operations. Additionally, image transformations are conducted using torchvision's transforms module. Our computational tasks are executed on Google Colab Pro+, which features a GPU A100 accelerator. This setup offers exceptional performance, driven by an NVIDIA A100 GPU with 6,912 CUDA cores and 40 GB of high-bandwidth memory, alongside an Intel Xeon processor and SSD-based storage.

Dataset Split In the proposed DCRUNet++ model, the dataset undergoes a stratified split into training (70%), testing (10%), and validation (20%) subsets. Initially, 10% of the data is allocated to the testing set, which is then evenly divided into distinct testing and validation subsets. This method ensures a balanced distribution of data across all three subsets, enhancing the robustness of model evaluation.

Table 1. Hyperparameters used for training the model.

Input image size	256x256x3
Number of epochs	100
Batch size	2
Patience	6
Learning rate	3×10^{-4}
Optimizer	Adam

3.3 Results

To rigorously evaluate the effectiveness of the proposed DCRUNet++ model for brain tumor segmentation, we performed a series of experiments on the TCGA-LGG dataset, comprising a The hyperparameters, critical for the model’s training and inference processes, were meticulously configured according to the detailed specifications outlined in Table 1, including the learning rate, batch size, number of epochs, optimizer settings, data augmentation techniques, loss functions, and regularization methods. The performance metrics obtained from our evaluations are as follows: Mean IoU of 0.9155, Dice Coefficient of 0.9467, Mean Accuracy of 0.9991, and Loss of 0.0708. These results were visualized through respective plots, demonstrating the model’s performance across various evaluation metrics. comprehensive collection of lower-grade glioma images.

The accuracy plot for the proposed DCRUNet++ model for brain tumor segmentation is shown in Fig. 4(a). The Dice Coefficient, with a value close to 1, confirms the model’s proficiency in accurate segmentation, maintaining a high level of overlap while minimizing false positives and negatives. The near-perfect Mean Accuracy suggests the model’s predictions are highly consistent with the ground truth across all pixels, underscoring its robustness and generalization capability, as shown in Fig. 4(b). The low loss value indicates minimal discrepancy between predicted outputs and actual labels during training, suggesting optimal parameter tuning, as shown in Fig. 4(c). The high Mean IoU indicates substantial overlap between predicted segmentation and ground truth, reflecting precise boundary delineation capabilities, as shown in Fig. 4(d). Several segmented images are shown in Fig. 5. The leftmost images are the original images, the middle ones are the original masks, and the rightmost are the segmented

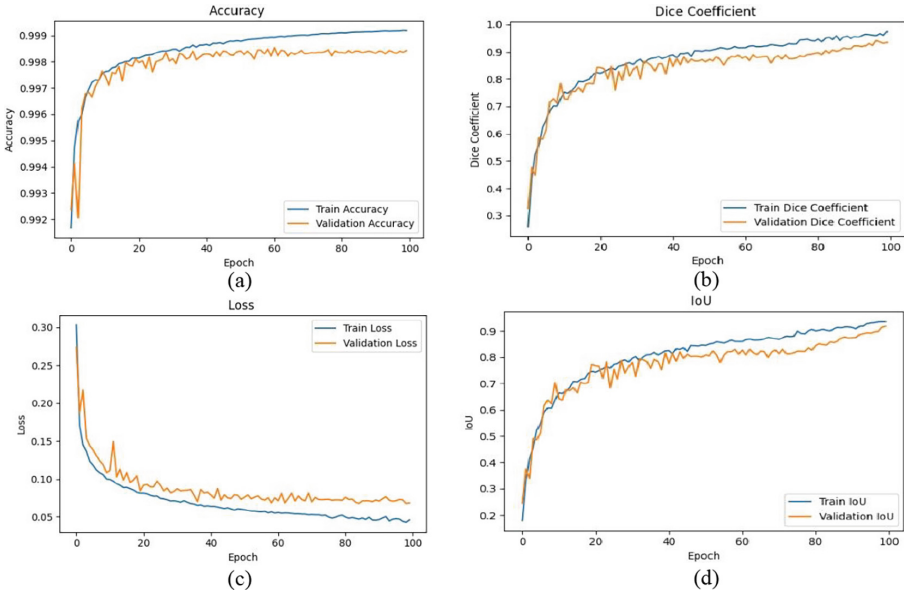


Fig. 4. Training and validation curves of (a) Accuracy, (b) Dice Coefficient, (c) Loss, and (d) IoU

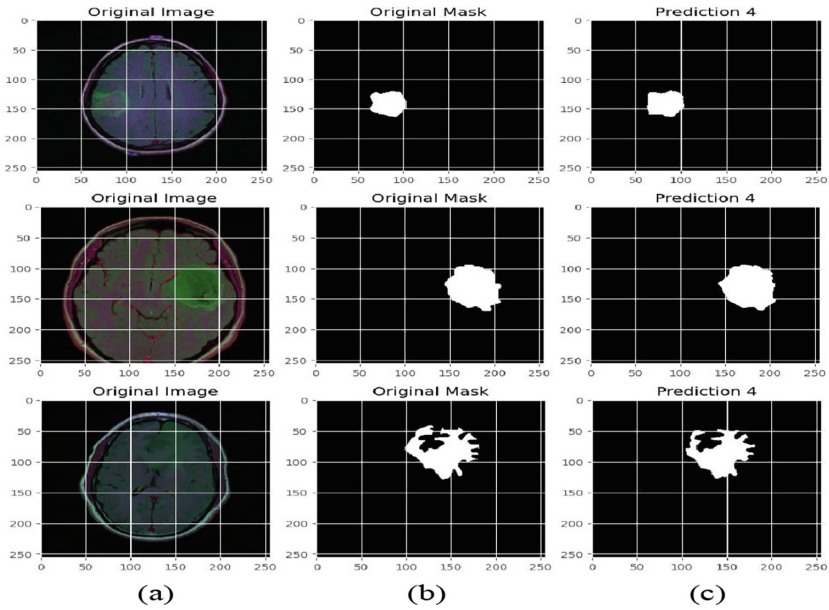


Fig. 5. (a) Independent Image, (b) Ground Truth Image, and (c) Segmented Image by DCRUNet++ Model

images by our proposed DCRUNet++ model. As we can clearly see in Fig. 5, the segmented images by the proposed DCRUNet++ model closely match the original masks in most cases, showing exact boundaries as the original tumor.

The superior performance of the DCRUNet++ model can be attributed to several key architectural and methodological innovations: novel RMs enhance gradient flow and mitigate the vanishing gradient problem, enabling the training of deeper networks for robust feature learning; Dense Skip Connections aggregate features from all preceding layers, ensuring effective utilization of multi-scale information throughout the network; Learnable Upsampling with Conv2D Transpose Layers optimizes spatial dimension increases, preserving important details and enhancing segmentation quality; SE Blocks recalibrate channel-wise feature responses, improving the model's focus on tumor regions and boosting accuracy; DS, incorporating multiple loss functions at various intermediate layers, compels the network to produce valuable features at multiple scales and depths, enhancing overall learning and performance; Comprehensive Data Augmentation and Regularization, including extensive data augmentation techniques and regularization methods like dropout and weight decay, increase training data diversity and prevent overfitting, ensuring robust performance on test data. Collectively, these results validate the DCRUNet++ model's efficacy and robustness, underscoring its potential for clinical application in aiding the diagnosis and treatment planning for glioma patients.

3.4 Comparative Analysis

In recent studies, various approaches have been employed to tackle brain tumor segmentation using MRI images. Buda et al. [3] utilized the FLAIR dataset from Kaggle LGG and reported an average Dice coefficient of 0.82, indicating a reasonably good performance. Similarly, Naser et al. [24] achieved a Dice coefficient of 0.84 in their segmentation task. Building on these efforts, Kumar et al. [17] further advanced the field by addressing the challenge of differing intensities and merging boundaries between brain tissues and tumor regions. Their model, trained on the same dataset, attained impressive segmentation results, boasting a Dice coefficient of 0.9056 and mean IoU of 0.8293. To the best of our knowledge, these are the only three papers that have been published on the Kaggle LGG dataset. In our research, we aimed to surpass these benchmarks by developing a novel DCRUNet++ model for brain tumor segmentation using the FLAIR image dataset. The proposed DCRUNet++ model integrates advanced architectural features such as RMs, dense skip connections, learnable upsampling with Conv2D transpose layers, and SE blocks, which collectively enhance feature extraction, gradient flow, and spatial resolution retention. These innovations contribute to the model's superior performance. The proposed DCRUNet++ model achieved a Dice coefficient of 0.9467 and an IoU of 0.9155, significantly outperforming previous studies and demonstrating competitive performance, as shown in Table 2. This substantial improvement underscores the efficacy of our model in accurately segmenting brain tumors, providing a robust tool for clinical applications in MRI-based diagnostics.

Table 2. Comparison of the Proposed DCRUNet++ Model with Existing Models

Author	Modalities	Dataset	IoU	Dice Coefficient	Accuracy
Buda et al. [3]	FLAIR	TCGA LGG	NA	0.82	NA
Naser et al. [24]	FLAIR	TCGA LGG	NA	0.84	0.9200
Chakroborty et al. [4]	FLAIR	TCGA LGG	0.8765	0.9056	0.9984
Kumar et al. [17]	FLAIR	TCGA LGG	0.8293	0.9056	0.9956
Kunjumon et al. [18]	FLAIR	TCGA LGG	0.8300	0.9230	0.9980
Kamal et al. [10]	FLAIR	TCGA LGG	0.8900	NA	0.9600
Proposed DCRUNet++	FLAIR	TCGA LGG	0.9155	0.9467	0.9991

4 Conclusion and Future scope

In this research, the DCRUNet++ model was developed for brain tumor segmentation using MRI images from the TCGA-LGG dataset, featuring advanced architectural enhancements like RMs, dense skip connections, learnable upsampling with Conv2D transpose layers, and SE blocks, achieving a Dice coefficient of 0.9467 and an IoU of 0.9155, which significantly outperform previous models and demonstrate its robustness and efficacy for clinical applications. The high accuracy and low loss observed during testing underscore the model's effectiveness in accurately segmenting brain tumors. Our approach demonstrates substantial improvements over existing methods by addressing challenges such as differing intensities and merging boundaries between brain tissues and tumor regions. However, the model's increased computational complexity due to custom loss functions, dense skip connections, and residual modules, coupled with its sensitivity to hyperparameter tuning and lack of external validation on independent datasets, limits its practical implementation in resource-constrained environments and raises concerns about its generalizability and reliability across diverse clinical settings, with future work focusing on refinement, exploration of additional datasets, and integration into clinical workflows.

References

1. Aishwarja, A.I., Eva, N.J., Mushtary, S., Tasnim, Z., Khan, N.I., Islam, M.N.: Exploring the machine learning algorithms to find the best features for predicting the breast cancer and its recurrence. In: Intelligent Computing and Optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020). pp. 546–558. Springer (2021)
2. Anaya-Isaza, A., Mera-Jiménez, L.: Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging. *IEEE Access* **10**, 23217–23233 (2022)
3. Buda, M., Saha, A., Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Comput. Biol. Med.* **109**, 218–225 (2019)

4. Chakroborty, P., Mishu, S.Z., Al Mamun, M., Hossain, M.A., Srizon, A.Y.: Predicting brain tumor region from mri flair images using ensemble method. In: 2023 26th International Conference on Computer and Information Technology (ICCIT). pp. 1–7. IEEE (2023)
5. Cinar, N., Ozcan, A., Kaya, M.: A hybrid densenet121-unet model for brain tumor segmentation from mr images. *Biomed. Signal Process. Control* **76**, 103647 (2022)
6. Dash, D.P., Kolekar, M., Chakroborty, C., Khosravi, M.R.: Review of machine and deep learning techniques in epileptic seizure detection using physiological signals and sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing* **23**(1), 1–29 (2024)
7. Farheen, F., Shamil, M.S., Ibtehaz, N., Rahman, M.S.: Segmentation of lung tumor from ct images using deep supervision. arXiv preprint [arXiv:2111.09262](https://arxiv.org/abs/2111.09262) (2021)
8. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. *Comput. Vis. Image Underst.* **114**(6), 712–722 (2010)
9. Habuza, T., Navaz, A.N., Hashim, F., Alnajjar, F., Zaki, N., Serhani, M.A., Statsenko, Y.: Ai applications in robotics, diagnostic image analysis and precision medicine: Current limitations, future trends, guidelines on cad systems for medicine. *Informatics in Medicine Unlocked* **24**, 100596 (2021)
10. Halloum, K., Ez-Zahraouy, H.: Advancing brain tumour segmentation: A novel cnn approach with resnet50 and drvu-net: A comparative study. *Intelligent Decision Technologies (Preprint)*, 1–18
11. 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)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
13. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. IEEE (2020)
14. Kanjo, E., Younis, E.M., Ang, C.S.: Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion* **49**, 46–56 (2019)
15. Kolekar, M.H., Bose, S., Pai, A.: Sarain-gan: Spatial attention residual unet based conditional generative adversarial network for rain streak removal. *IEEE Access* (2024)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
17. Kumar, P.S., Sakthivel, V., Raju, M., Sathya, P.: Brain tumor segmentation of the flair mri images using novel resunet. *Biomed. Signal Process. Control* **82**, 104586 (2023)
18. Kunjumon, A., Jacob, C., Resmi, R.: An efficient u-net based model for low grade glioma segmentation in mri images. In: 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE). pp. 1–5. IEEE (2024)
19. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial intelligence and statistics. pp. 562–570. Pmlr (2015)

20. Li, C., Song, X., Zhao, H., Feng, L., Hu, T., Zhang, Y., Jiang, J., Wang, J., Xiang, J., Sun, Y.: An 8-layer residual u-net with deep supervision for segmentation of the left ventricle in cardiac ct angiography. *Comput. Methods Programs Biomed.* **200**, 105876 (2021)
21. Liu, P., Dou, Q., Wang, Q., Heng, P.A.: An encoder-decoder neural network with 3d squeeze-and-excitation and deep supervision for brain tumor segmentation. *IEEE Access* **8**, 34029–34037 (2020)
22. Menze, B., Isensee, F., Wiest, R., Wiestler, B., Maier-Hein, K., Reyes, M., Bakas, S.: Analyzing magnetic resonance imaging data from glioma patients using deep learning. *Comput. Med. Imaging Graph.* **88**, 101828 (2021)
23. Metlek, S., Çetiner, H.: Resunet+: A new convolutional and attention block-based approach for brain tumor segmentation. *IEEE Access* (2023)
24. Naser, M.A., Deen, M.J.: Brain tumor segmentation and grading of lower-grade glioma using deep learning in mri images. *Comput. Biol. Med.* **121**, 103758 (2020)
25. Ruba, T., Tamilselvi, R., Beham, M.P.: Brain tumor segmentation in multimodal mri images using novel lsis operator and deep learning. *J. Ambient. Intell. Humaniz. Comput.* **14**(10), 13163–13177 (2023)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)* (2014)
27. Singh, V.K., Kolekar, M.H.: Deep learning empowered covid-19 diagnosis using chest ct scan images for collaborative edge-cloud computing platform. *Multimedia Tools and Applications* **81**(1), 3–30 (2022)
28. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 31 (2017)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1–9 (2015)
30. Walsh, J., Othmani, A., Jain, M., Dev, S.: Using u-net network for efficient brain tumor segmentation in mri images. *Healthcare Analytics* **2**, 100098 (2022)
31. Wang, L., Lee, C.Y., Tu, Z., Lazebnik, S.: Training deeper convolutional networks with deep supervision. *arXiv preprint [arXiv:1505.02496](https://arxiv.org/abs/1505.02496)* (2015)
32. Yadav, A.C., Kolekar, M.H., Zope, M.K.: Resnet-101 empowered deep learning for breast cancer ultrasound image classification. In: *BIOSTEC* (1). pp. 763–769 (2024)
33. 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)



DrowzEE-G-Mamba: Leveraging EEG and State Space Models for Driver Drowsiness Detection

Gourav Siddhad^(✉), Sayantan Dey, and Partha Pratim Roy

Department of Computer Science and Engineering, Indian Institute of Technology,
Roorkee, Roorkee, 247667, Uttarakhand, India
{g_siddhad,partha}@cs.iitr.ac.in, sayantan.cs@srict.iitr.ac.in

Abstract. Driver drowsiness is identified as a critical factor in road accidents, necessitating robust detection systems to enhance road safety. This study proposes a driver drowsiness detection system, DrowzEE-G-Mamba, that combines Electroencephalography (EEG) with State Space Models (SSMs). EEG data, known for its sensitivity to alertness, is used to model driver state transitions between alert and drowsy. Compared to traditional methods, DrowzEE-G-Mamba achieves significantly improved detection rates and reduced false positives. Notably, it achieves a peak accuracy of 83.24% on the SEED-VIG dataset, surpassing existing techniques. The system maintains high accuracy across varying complexities, making it suitable for real-time applications with limited resources. This robustness is attributed to the combination of channel-split, channel-concatenation, and channel-shuffle operations within the architecture, optimizing information flow from EEG data. Additionally, the integration of convolutional layers and SSMs facilitates comprehensive analysis, capturing both local features and long-range dependencies in the EEG signals. These findings suggest the potential of DrowzEE-G-Mamba for enhancing road safety through accurate drowsiness detection. It also paves the way for developing powerful SSM-based AI algorithms in Brain-Computer Interface applications.

Keywords: Cognitive State Monitoring · Driver Fatigue · EEG · Mamba · Safety · State Space Model

1 Introduction

Driver drowsiness detection is crucial for road safety, as fatigue and sleepiness are major causes of car crashes, often leading to severe injuries or fatalities. Unlike intoxication, drowsiness develops gradually and can be unnoticed by drivers. Effective detection systems can prevent accidents by alerting drivers to take corrective actions, such as resting. With the rise of advanced driver-assistance systems (ADAS) [28] and autonomous vehicles, integrating robust drowsiness detection is essential for enhancing transportation safety and reliability. These

systems not only protect individual drivers but also contribute to public safety by reducing drowsiness-induced accidents.

EEG is a valuable tool for real-time detection and analysis of cognitive states, capturing the brain's electrical activity [34]. EEG measures voltage fluctuations from neuronal ionic currents, offering insights into mental states like attention, alertness, fatigue, and cognitive load [29]. Its high temporal resolution is ideal for monitoring rapid changes in brain activity, making it perfect for transient cognitive state monitoring. By analyzing frequency bands (delta, theta, alpha, beta, and gamma) and spatial distribution, researchers can infer neural mechanisms behind various cognitive processes [26]. This capability is crucial in brain-computer interfaces (BCIs), neurofeedback, and cognitive neuroscience research. EEG's non-invasive nature and relatively low cost enhance its practicality for cognitive state detection, advancing both clinical and real-world applications.

Due to the complex, non-linear nature of EEG data, standard deep learning models struggle with accurate analysis. This study explores Mamba [14], a state-of-the-art state-space model (SSM), for effective driver drowsiness detection using EEG signals. Mamba excels at capturing the intricate patterns and non-linearities within EEG data. It extracts relevant features and integrates them with a hidden state space, reflecting the underlying brain activity. This allows Mamba to effectively manage noise and uncertainties inherent in EEG data, leading to more accurate drowsiness detection. Additionally, Mamba's efficient feature extraction and adaptive learning capabilities make it ideal for real-time monitoring and prediction, surpassing traditional EEG-based methods. Building upon the advantages of structured SSMs [14], Mamba offers computational efficiency and excels at capturing long-range dependencies within data. Notably, Mamba addresses limitations of previous models by incorporating time-varying parameters and employing a novel hardware-aware algorithm for efficient training and inference [44]. This versatility has been demonstrated in various visual tasks, including ImageNet classification [44], remote sensing image classification [5], image dehazing [43], point cloud analysis [21], and medical image segmentation [31], showcasing Mamba's potential beyond driver drowsiness detection and opening new avenues for research in computational neuroscience.

This paper introduces a driver drowsiness detection system using EEG data and the Mamba state-space model. Mamba's ability to handle complex brain activity dynamics makes it ideal for analyzing drowsiness-related EEG changes. The system leverages Mamba's robustness and adaptability to noise and non-linearity in EEG signals. This Mamba-based approach aims to surpass existing methods by providing a more precise and responsive solution, potentially reducing fatigue-related accidents. Additionally, Mamba's advanced feature extraction capabilities offer broader applications in computational neuroscience and BCIs, as demonstrated by its effectiveness in distinguishing cognitive loads. Integrating Mamba into EEG research holds promise for unlocking new discoveries in brain function. The key contributions of this work are as follows:

- This research introduces DrowzEE-G-Mamba, a novel deep learning model leveraging the Mamba state-space model for real-time driver drowsiness detec-

tion using EEG data. DrowzEE-G-Mamba surpasses existing methods by achieving a peak accuracy of 83.24% on the SEED-VIG dataset.

- DrowzEE-G-Mamba demonstrates exceptional robustness, maintaining high accuracy across varying model complexities. Notably, it achieves a remarkable 83.24% accuracy even with a minimal 10.1k parameters. This efficiency translates to faster training, lower memory footprint, and easier deployment on resource-constrained devices.
- DrowzEE-G-Mamba exhibits a smaller confidence interval compared to other methods, indicating greater consistency in performance. This, coupled with its adaptability across various computational settings, suggests its potential for diverse practical applications beyond driver drowsiness detection, opening doors for real-time brain activity monitoring in other domains.

This paper presents a methodical exploration of EEG-based fatigue detection and its potential for enhancing road safety technologies. In Section 2, a review of recent literature on driver drowsiness and vigilance is conducted. Section 3 details the methodology employed in this research. The empirical findings of the study are presented in Section 4. Finally, the discussion in Section 5 extends beyond the results, exploring the broader implications and future directions for this research.

2 Related Work

Early research in EEG-based fatigue detection identified biomarkers such as variations in theta and alpha EEG frequency bands [15]. Deep learning has further transformed EEG analysis, with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) adeptly handling spatial and temporal data [32]. Hybrid models combining CNNs with RNNs or other techniques enhance feature extraction [3], offering superior accuracy and computational efficiency for real-time applications [38]. These models manage large, complex datasets without extensive feature engineering, outperforming traditional methods [37]. However, variability in EEG signals across individuals affects model generalization [17].

Driver drowsiness detection utilizes physiological (EEG, ECG, EOG) [18], vehicle behavior (steering, lane departure, pedal use) [9], and behavioral (facial expressions, head position, eye closure) [4] signals to assess driver state. Physiological methods are accurate but intrusive, while vehicle-based and behavioral methods offer non-intrusive detection but may be less accurate. Recent advancements integrate multiple detection methods (physiological, behavioral, vehicle-based) for improved drowsiness detection accuracy and reliability. Real-world EEG systems face challenges: discomfort from traditional setups, artifact vulnerability, and inter-individual variability requiring personalized models [11]. Future systems should prioritize comfort (dry electrodes, wireless headsets), robust artifact removal, and real-time processing with efficient algorithms. CNNs effectively extract features from EEG signals [2], and Transformers excel at handling time-series data and capturing long-range dependencies in EEG for tasks like mental state classification and seizure detection [33, 36].

State space models (SSMs) offer a powerful tool in neuroscience to decipher complex neural dynamics and behaviors. These models describe systems evolving over time, inferring hidden states and underlying processes from observed neural data [31]. This allows researchers to gain insights into neural activity, dynamics, and behavior. SSMs are particularly useful for decoding neural activity to infer hidden cognitive states, illustrating how neural populations interact and evolve, and linking neural activity with behavior. A prominent application lies in brain-machine interfaces (BMIs). For example, Wu et al. [39] used a Kalman filter (an SSM) for real-time motor cortex decoding. Churchland et al. [6] analyzed motor cortex dynamics with SSMs. Mante et al. [24] studied decision-making in the prefrontal cortex using SSMs. Despite these advantages, such as flexibility for diverse data types, hidden state inference, and prior knowledge integration, challenges remain. These include computational intensity, high-quality data requirements, and difficulty interpreting the biological relevance of inferred hidden states. Future research may focus on improving computational methods, integrating multimodal data, and enhancing model interpretability.

While initially limited by computational demands, SSMs have evolved. The Structured State Space Sequence Model (S4) [13] addresses this with efficient kernel computations. Additionally, SSMs are now integrated into various deep learning architectures [35]. However, constant sequence transformation restricts context-based reasoning in standard models. Recent advancements like Mamba (Selective SSM) introduce time-varying parameters for more efficient training and inference [12]. This paves the way for applying SSMs to computer vision tasks, similar to Transformers in NLP. Studies like ViS4mer [16] and S4ND [27] utilize SSM blocks for modeling visual data across dimensions. VMamba [22] and Vim [44] address direction-sensitivity and global context modeling, respectively. SSMs are a powerful framework in neuroscience, providing deep insights into neural dynamics and behavior. They decode neural activity, model population dynamics, and study cognitive processes. As computational techniques and data quality improve, SSMs are likely to play an even more critical role in advancing our understanding of the brain.

3 Methodology

This section examines the foundational concepts underlying DrowzEE-G-Mamba, a deep learning model designed for driver drowsiness detection using EEG data. These concepts, such as State Space Models (SSMs) and their discretization process, are essential for capturing the complex relationships within EEG signals. DrowzEE-G-Mamba's overall architecture is then discussed which is adapted from MedMamba [40]. 2D-Selective-Scan mechanism, adapted from VMamba [22], is highlighted as crucial for extracting informative features from the EEG data. Finally, the detailed modeling process of the SS-Conv-SSM block, the fundamental building block of DrowzEE-G-Mamba, is examined to understand how features indicative of drowsiness are efficiently extracted from EEG signals.

3.1 Preliminaries

Recent SSM-based models, such as Structured State Space Sequence Models (S4) and Mamba, utilize a classical continuous system to map a 1D input function or sequence, denoted as $x(t) \in \mathcal{R}$, through intermediate implicit states $h(t) \in \mathcal{R}^N$, to an output $y(t) \in \mathcal{R}$. This process can be represented by a linear Ordinary Differential Equation (ODE) [12, 22]:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \quad (1)$$

Here, $\mathbf{A} \in \mathcal{R}^{N \times N}$ represents the state matrix, while $\mathbf{B} \in \mathcal{R}^{N \times 1}$ and $\mathbf{C} \in \mathcal{R}^{N \times 1}$ denote the projection parameters.

The S4 Model and Mamba leverage discretization to make continuous systems compatible with deep learning architectures. This process introduces a timescale parameter, denoted by Δ , which transforms the continuous system matrices \mathbf{A} and \mathbf{B} into their discrete counterparts, denoted by $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$. A common discretization rule employed for this purpose is the zero-order hold (ZOH).

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \end{aligned} \quad (2)$$

After applying discretization with a step size Δ , Equation 1 transforms into a linear recurrence form (Equation 3) as follows:

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \quad (3)$$

This equation represents the state update (h') based on the previous state (h) and the current input (x). Additionally, the output (y) is obtained by multiplying the current state with an output matrix (\mathbf{C}).

Finally, the SSM model employs a global convolution to efficiently capture long-range dependencies within the input sequence:

$$\begin{aligned} \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}\overline{\mathbf{B}}, \dots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}) \\ y &= x * \overline{\mathbf{K}} \end{aligned} \quad (4)$$

This convolution utilizes a structured kernel ($\overline{\mathbf{K}}$), which incorporates the discretized state transition matrices ($\overline{\mathbf{A}}, \overline{\mathbf{B}}$) and the output matrix (\mathbf{C}). The length of the input sequence x is denoted by L .

3.2 DrowzEE-G-Mamba Architecture

DrowzEE-G-Mamba is a deep learning model proposed for driver drowsiness detection. It takes inspiration from the architectural design and concepts of Med-Mamba and VMamba. It utilizes a patch embedding layer to convert raw EEG

data into a format suitable for subsequent processing. The model’s core consists of stacked SS-Conv-SSM blocks, to capture complex spatio-temporal features within EEG signals indicative of drowsiness. Patch merging layers downsample the extracted features, facilitating efficient processing and classification. Finally, a feature classifier accurately identifies drowsiness states based on the learned feature representations.

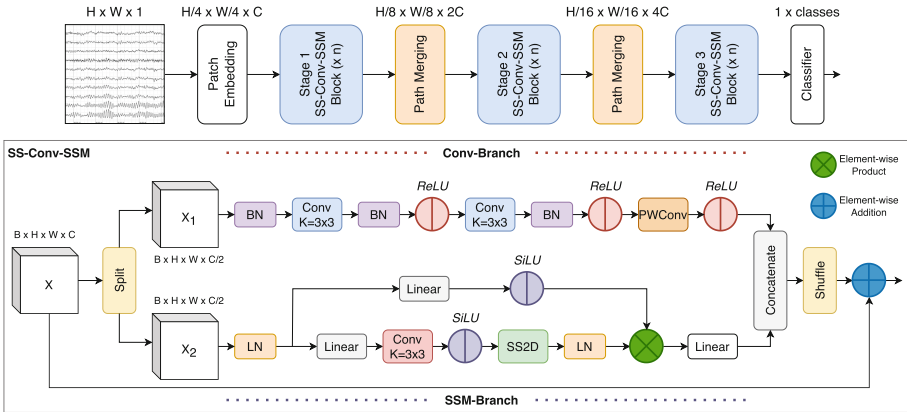


Fig. 1. Architecture of DrowzEE-G-Mamba: BN, LN, linear, PWConv, and DWConv represent batch normalization, layer normalization, linear layer, point-wise convolution, and depth-wise convolution, respectively.

Figure 1 illustrates the DrowzEE-G-Mamba model architecture, which processes EEG data in a series of multiple stacked stages. The model begins by transforming the raw EEG data (dimensions $H \times W \times 1$) into a format suitable for subsequent processing through a patch embedding layer. The data then undergoes a series of processing stages, each consisting of multiple SS-Conv-SSM blocks followed by patch merging operations. These stages progressively reduce the spatial dimensions of the feature maps while increasing the channel dimensions. The final output of this processing pipeline is fed into a classifier, which predicts the driver’s drowsiness state.

The core of DrowzEE-G-Mamba lies in its stacked SS-Conv-SSM blocks (detailed structure in Figure 1 bottom section). These blocks are specifically designed to capture the intricate spatio-temporal features within EEG signals that are crucial for drowsiness detection. Each block consists of two branches: Conv-Branch and SSM-Branch. Conv-Branch focuses on extracting local features through standard operations like batch normalization (BN), convolutions (Conv), pointwise convolutions (PWConv), and ReLU activations. SSM-Branch leverages linear layers, depth-wise convolutions (DWConv), SiLU activations, and structured state space 2D (SS2D) components to capture long-range dependencies and global context within the EEG data. Finally, element-wise addition and concatenation operations combine the features from both branches.

A key aspect is the inclusion of a shuffle operation at the end of the block. This helps mitigate potential information loss caused by the initial channel split within the SS-Conv-SSM architecture. This dual-branch design empowers DrowzEE-G-Mamba to efficiently learn complex patterns from EEG data, making it well-suited for driver drowsiness detection and other cognitive state analysis tasks. Inspired by ViTs, DrowzEE-G-Mamba employs a patch embedding layer as the first processing step. This layer transforms the raw EEG data, denoted as $x \in R^{H \times W \times 1}$, into non-overlapping patches of size 4×4 . The patch embedding layer achieves this transformation by mapping the single channel dimension to a higher dimensionality (C) without flattening the EEG data into a one-dimensional sequence. This approach preserves the two-dimensional (2D) structure of the EEG data, which is crucial for capturing spatial relationships within the signals. As a result, the patch embedding layer generates a feature map with dimensions $\frac{H}{4} \times \frac{W}{4} \times C$.

Following the patch embedding, DrowzEE-G-Mamba leverages stacked SS-Conv-SSM blocks in Stage 1 to process the feature map. These blocks are designed to extract informative features from the EEG data. Crucially, they capture both local details and long-range dependencies within the signals. Importantly, the dimensions of the feature map remain unchanged in this stage, allowing the model to focus on extracting rich features without altering the spatial resolution. To create hierarchical representations of the EEG data, patch merging layers are employed after Stage 1. These layers perform down-sampling, progressively reducing the spatial resolution (denoted by H and W) of the feature maps. In contrast, the channel dimension (denoted by C) typically doubles after each patch merging layer. Stages 2, 3, and 4 repeat this process, resulting in progressively lower spatial resolutions (e.g., $\frac{H}{16} \times \frac{W}{16} \times 4C$ for Stage 2) and increased channel dimensions. This down-sampling allows the model to learn complex patterns across different scales of the EEG data while maintaining computational efficiency. At the end of the network, a classifier with an adaptive global pooling layer and a linear layer determines the category of the input.

3.3 2D Selective Scan

The 2D-selective-scan (SS2D) proposed by VMamba, is a core element of Med-Mamba. SS2D adapts the selective scan space state sequence model (S6) designed for natural language processing to address the “direction-sensitive” problem in S6. To bridge the gap between 1-D array scanning and 2-D plane traversing, SS2D introduces a Cross-Scan Module (CSM). CSM uses a four-way scanning strategy, scanning from four corners across the feature map to the opposite locations, ensuring each pixel integrates information from all directions, achieving a global receptive field without increasing computational complexity.

By incorporating CSM, SS2D maintains the linear complexity of S6 while capturing long-range dependencies, essential for accurate medical image classification. SS2D comprises three components: a scan expanding operation (CSM), an S6 block, and a scan merging operation. The scan expanding operation unfolds the input image along four directions (top-left to bottom-right, bottom-right to

top-left, top-right to bottom-left, and bottom-left to top-right) into sequences. The S6 block processes these sequences to extract features, ensuring thorough scanning from various directions. Finally, the four directional features are merged through scan merging to reconstruct the 2D feature map, resulting in an output of the same size as the input. The S6 block, derived from Mamba, introduces a selective mechanism based on S4 by adjusting SSM parameters according to input. This enables the model to distinguish and retain relevant information while filtering out irrelevant details. The detailed pseudo-code for the S6 block can be found in the MedMamba [40].

3.4 SS-Conv-SSM Block

A hybrid basic block named SS-Conv-SSM, utilized in this work was introduced in MedMamba [40]. This block integrates convolutional layers for extracting local features with SSM’s ability to capture long-range dependencies. A grouped convolution, introduced in AlexNet [19], uses multiple kernels per layer to promote learning various high and low level features was also incorporated into the SS-Conv-SSM. SS-Conv-SSM is a lightweight dual-branch block (Figure 1). It partitions the feature map into two groups using channel-split, then extracts global and local information from each group through the Conv-Branch and SSM-Branch, respectively. Finally, channel-concatenation restores the channel dimension size, and channel-shuffle ensures information is not lost between channels due to grouped convolution operations [41]. Following the settings of classic CNNs and ViTs, the activation functions in the Conv-Branch and SSM-Branch are set to ReLU [1] and SiLU [10], respectively.

The modeling process of SS-Conv-SSM for feature maps is formalized. Given a module input $x \in R^{H \times W \times C}$ and a module output $y \in R^{H \times W \times C}$, f is used to represent the channel-split, and then there is

$$x \in R^{H \times W \times C} \quad x_{i=1,2} \in R^{H \times W \times \frac{C}{2}}$$

Next, the f^{-1} and g are used to represent channel-concatenation and channel-shuffle respectively. To match the convolution operation, a permute operation is utilized to rearrange the original feature map. Based on the above, the modeling process of Conv-Branch can be defined as follows:

$$\begin{aligned} \bar{x}_1 &\in R^{\frac{C}{2} \times H \times W} \leftarrow \text{permute}(x_1) \\ x_1' &= \text{BatchNorm}_1(\bar{x}_1) \\ x_1'' &= \text{ReLU}(\text{BatchNorm}_2(\text{Conv}_{3 \times 3}(x_1'))) \\ x_1''' &= \text{ReLU}(\text{BatchNorm}_3(\text{Conv}_{3 \times 3}(x_1''))) \\ \widehat{x}_1 &= \text{ReLU}(\text{PWConv}(x_1''')) \\ \widetilde{x}_1 &\in R^{H \times W \times \frac{C}{2}} \leftarrow \text{permute}(\widehat{x}_1) \end{aligned}$$

Meanwhile, the modeling process of SSM-Branch can be defined as follows:

$$\begin{aligned}\bar{x}_2 &= LayerNorm_1(x_2) \\ x_2' &= SiLU(DWConv(Linear(\bar{x}_2))) \\ x_2'' &= LayerNorm_2(SS2D(x_2')) \\ x_2''' &= SiLU(Linear(\bar{x}_2)) \\ \widetilde{x}_2 &= Linear(x_2'' \otimes x_2''')\end{aligned}$$

In summary, the output of SS-Conv-SSM be formulated as follows:

$$y = x \oplus g(f^{-1}(\widetilde{x}_1, \widetilde{x}_2))$$

4 Results and Discussion

This section presents the findings of this study and analyzes their significance for the field of driver drowsiness research. The analysis focuses on the effectiveness of the employed methods and the implications of the observed outcomes. This is followed by a comparative analysis with relevant findings from existing literature to contextualize our results.

4.1 Experimental Data

This study utilizes the SEED-VIG dataset [42], a valuable open-source resource designed to investigate driver vigilance and drowsiness through EEG recordings. The dataset offers a diverse subject pool, encompassing recordings from 23 participants. To enhance real-world applicability, participants engaged in a driving simulation designed to closely mimic real-world driving conditions. EEG recordings were captured using a 17-channel montage based on the international 10-20 system. This montage specifically targeted key temporal (FT7, FT8, T7, T8, TP7, TP8) and posterior (CP1, CP2, P1, PZ, P2, PO3, POZ, PO4, O1, OZ, O2) brain regions, ensuring comprehensive coverage of brain activity relevant to vigilance and drowsiness. High temporal resolution, crucial for detailed analysis, was achieved with a sampling rate of 1000 Hz. Sessions were strategically scheduled post-lunch to encourage the onset of fatigue in participants.

Drowsiness states were quantified using the PERCLOS (percentage of eyelid closure) metric. A threshold of 0.5 was employed to classify PERCLOS values into “awake” and “drowsy” states, enabling a binary classification approach for evaluating driver fatigue detection methods. To minimize artifacts and improve computational efficiency, EEG signals were band-pass filtered (1-75 Hz) and down-sampled to 200 Hz. Subsequently, the data was segmented into one-second epochs, resulting in a standardized format of (17, 200, 1) per epoch. The entire dataset comprised approximately 40,710 samples and was divided into training (70%), validation (15%), and test (15%) sets to facilitate model development and evaluation.

4.2 Implementation Details

The computational environment consisted of a DELL Precision 7820 Tower Workstation equipped with Ubuntu 22.04 operating system, an Intel Core(TM) Xeon Silver 4216 CPU, and an NVIDIA RTX A4000 12GB GPU. This hardware configuration facilitated the implementation of Deep Learning (DL) models using Python 3.12 and the PyTorch library. The Adam optimizer, recognized for its efficiency, was employed with its default hyperparameters ($\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Both EEGNet and TSception models underwent training for 100 epochs, utilizing a batch size of 16 and a learning rate of $1e - 4$. For the Support Vector Machine (SVM) classification, the Radial Basis Function (RBF) kernel from scikit-learn [30] was implemented with its default settings. Stratified five-fold cross-validation was employed to assess classification accuracy, with the results averaged for a robust evaluation.

4.3 Classifiers

This work employs a balanced evaluation approach using three established classifiers for EEG-based emotion classification. Support Vector Machine (SVM) [7] is a popular supervised learning model for classification, known for its ability to maximize the class margin for new data points. SVMs can handle non-linear classification through the kernel trick, effectively mapping inputs into high-dimensional spaces. EEGNet [20] is a CNN-based architecture that achieves competitive accuracy using deep and separable convolutions. It incorporates temporal convolution for learning frequency filters, depth-wise convolution for frequency-specific spatial filters, and separable convolution for efficient feature map combinations. TSception [8] utilizes a dynamic temporal layer to learn temporal and frequency representations from EEG channels. It also includes an asymmetric spatial layer for capturing global spatial patterns and emotional asymmetry, a high-level fusion layer, and a final classifier that leverages various convolutional kernel sizes for spatial analysis. ConvNext [23] is a state-of-the-art CNN architecture that achieves competitive performance on various image classification benchmarks. It incorporates design principles from recent transformer models to enhance feature learning and improve efficiency compared to traditional CNNs. LMDA-Net [25] is a lightweight deep learning model specifically designed for EEG-based emotion classification. It employs a multi-modal approach, combining temporal and spatial features, to effectively capture the complex patterns in EEG signals, resulting in efficient and accurate emotion recognition.

4.4 Evaluation

The results presented in Table 1 demonstrate the effectiveness of different methods for driver drowsiness detection on the SEED-VIG dataset. The evaluation revealed a clear hierarchy in the effectiveness of the compared methods for driver drowsiness detection on the SEED-VIG dataset. Support Vector Machine (SVM)

Table 1. Results of different methods on SEED-VIG dataset for driver drowsiness detection with 95% confidence interval

Method	Accuracy
SVM [7]	65.52 \pm 0.02
EEGNet [20]	80.74 \pm 0.75
TSception [8]	83.15 \pm 0.36
ConvNeXt [23]	81.95 \pm 0.61
LMDA-Net [25]	81.06 \pm 0.99
DrowzEE-G-Mamba	83.24 \pm 0.24

achieved the lowest accuracy (65.52%) with a narrow confidence interval (0.02), indicating consistent but limited performance. This suggests SVM may not adequately capture the complexities of EEG data for this task. EEGNet demonstrated a significant improvement over SVM, achieving an accuracy of 80.74%. However, its larger confidence interval (0.75) implies greater variability in performance. While superior to SVM, this suggests EEGNet might benefit from further optimization for drowsiness detection. TSception surpassed EEGNet with an accuracy of 83.15% and a reduced confidence interval (0.36), indicating both higher accuracy and more consistent performance. This suggests TSception’s architecture effectively captures relevant features in the EEG data. ConvNeXt achieved an accuracy of 81.95% and LMDA-Net obtained an accuracy of 81.06%. While their performance was comparable, DrowzEE-G-Mamba’s higher accuracy and lower variability make it a more reliable choice for real-time driver drowsiness detection.

DrowzEE-G-Mamba emerged as the most effective method, achieving the highest accuracy (83.24%) with the smallest confidence interval (0.24). This signifies not only superior detection accuracy but also the most consistent results. By combining EEG data with State Space Models (SSMs), DrowzEE-G-Mamba effectively models both local and long-range dependencies within the data, leading to superior drowsiness state detection. In conclusion, these findings highlight the clear advantage of DrowzEE-G-Mamba compared to traditional methods (SVM) and advanced neural network approaches (EEGNet, TSception, ConvNeXt, and LMDA-Net) for driver drowsiness detection on the SEED-VIG dataset. Its high accuracy and low variability make DrowzEE-G-Mamba a promising tool for real-time driver drowsiness detection, potentially contributing to accident prevention and improved road safety.

The chart in Figure 2 illustrates the accuracy of the DrowzEE-G-Mamba model on the SEED-VIG dataset for driver drowsiness detection, plotted against the number of parameters (in thousands). The model achieves an accuracy of 82.64% with 819k parameters. As the number of parameters decreases, the accuracy generally remains above 81%, with slight fluctuations. For instance, at 357k parameters, the accuracy is 82.11%, while at 209k parameters, it is 82.17%. The lowest number of parameters tested is 10.1k, where the accuracy maintains a

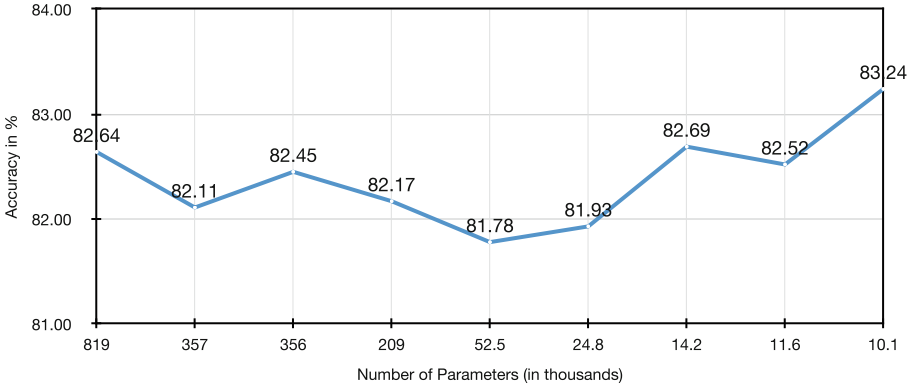


Fig. 2. This chart shows the impact of model complexity on driver drowsiness detection using the SEED-ViG dataset. It visualizes the relationship between average model accuracy (percentage) with 95% confidence interval and the number of parameters (thousands). As evident, accuracy increases with model complexity, ranging from 81.78% for an 819K parameter model to 83.24% for a 10.1K parameter model. One of the primary strategies employed to reduce the model size was the careful adjustment of hyper-parameters, specifically through the elimination of certain blocks within the model architecture.

robust peak at 83.24%. This chart demonstrates that DrowzEE-G-Mamba maintains high accuracy across a range of model complexities, with only a minor fluctuation in performance as the number of parameters decreases. The model which gave the highest accuracy with 10.1k parameters had one SS-Conv-SSM block with 32 dimensions.

The combined analysis of the presented table and chart suggests DrowzEE-G-Mamba’s exceptional potential as a highly effective and reliable model for driver drowsiness detection. Notably, the model achieves accuracy levels higher than the peak performance of leading models like TSception. Furthermore, DrowzEE-G-Mamba demonstrates a remarkable characteristic, it maintains this high accuracy across varying levels of model complexity (as reflected by different parameter counts). This robustness makes DrowzEE-G-Mamba particularly well-suited for real-time applications where computational resources might be constrained. The model’s consistency in performance is further emphasized by the narrow confidence interval and stable accuracy observed across parameter counts. This consistency underscores DrowzEE-G-Mamba’s suitability for practical deployment in real-world scenarios.

5 Conclusion

This research investigated the efficacy of DrowzEE-G-Mamba, a deep learning model for driver drowsiness detection using EEG data. DrowzEE-G-Mamba

achieved a peak accuracy of 83.24% on the SEED-VIG dataset, demonstrating its effectiveness. Notably, the model maintained high accuracy across varying parameter complexities, indicating strong robustness for real-time applications with limited computational resources. DrowzEE-G-Mamba's architecture balances sophistication with efficiency. The model leverages channel-split, channel-concatenation, and channel-shuffle operations to optimize information flow within the EEG data. DrowzEE-G-Mamba surpasses existing methods in two key aspects: accuracy and robustness. It achieves the highest accuracy while maintaining this performance even with a significant number of parameters. This translates to consistent and reliable detection, even with a larger computational footprint, making it a strong candidate for real-time driver drowsiness detection.

Overall, DrowzEE-G-Mamba presents a robust, efficient, and highly accurate solution for driver drowsiness detection. Its ability to function across diverse computational constraints makes it a promising tool for real-time drowsiness monitoring and enhancing road safety. Future work will focus on further optimization and explore applications in broader cognitive state detection tasks, expanding its impact and utility in various real-world scenarios. While challenges remain in refining accuracy and generalizability of fatigue detection systems, DrowzEE-G-Mamba's performance highlights the potential for significant advancements in real-time driver fatigue detection. Future research will target further accuracy improvements, applicability expansion, and integration into practical, real-world applications, ultimately contributing to safer driving environments.

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint [arXiv:1803.08375](https://arxiv.org/abs/1803.08375) (2018)
2. Aggarwal, S., Chugh, N.: Review of machine learning techniques for eeg based brain computer interface. *Archives of Computational Methods in Engineering* **29**(5), 3001–3020 (2022)
3. Ardabili, S.Z., Bahmani, S., Lahijan, L.Z., Khaleghi, N., Sheykhivand, S., Danishvar, S.: A novel approach for automatic detection of driver fatigue using eeg signals based on graph convolutional networks. *Sensors* **24**(2), 364 (2024)
4. Bergasa, L.M., Nuevo, J., Sotelo, M.A., Barea, R., Lopez, M.E.: Real-time system for monitoring driver vigilance. *IEEE Trans. Intell. Transp. Syst.* **7**(1), 63–77 (2006)
5. Chen, K., Chen, B., Liu, C., Li, W., Zou, Z., Shi, Z.: Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters* (2024)
6. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V.: Neural population dynamics during reaching. *Nature* **487**(7405), 51–56 (2012)
7. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995)
8. Ding, Y., Robinson, N., Zhang, S., Zeng, Q., Guan, C.: TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing* (2022)
9. Dong, Y., Hu, Z., Uchimura, K., Murayama, N.: Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Trans. Intell. Transp. Syst.* **12**(2), 596–614 (2010)

10. Elfving, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **107**, 3–11 (2018)
11. Fatourechhi, M., Bashashati, A., Ward, R.K., Birch, G.E.: Emg and eeg artifacts in brain computer interface systems: A survey. *Clin. Neurophysiol.* **118**(3), 480–494 (2007)
12. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint [arXiv:2312.00752](https://arxiv.org/abs/2312.00752) (2023)
13. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint [arXiv:2111.00396](https://arxiv.org/abs/2111.00396) (2021)
14. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural. Inf. Process. Syst.* **34**, 572–585 (2021)
15. Huang, R.S., Jung, T.P., Makeig, S.: Tonic changes in eeg power spectra during simulated driving. In: *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: 5th International Conference, FAC 2009 Held as Part of HCI International 2009 San Diego, CA, USA, July 19-24, 2009 Proceedings 5*. pp. 394–403. Springer (2009)
16. Islam, M.M., Bertasius, G.: Long movie clip classification with state-space video models. In: *European Conference on Computer Vision*. pp. 87–104. Springer (2022)
17. Kar, S., Bhagat, M., Routray, A.: Eeg signal analysis for the assessment and quantification of driver’s fatigue. *Transport. Res. F: Traffic Psychol. Behav.* **13**(5), 297–306 (2010)
18. Khushaba, R.N., Kodagoda, S., Lal, S., Dissanayake, G.: Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Trans. Biomed. Eng.* **58**(1), 121–131 (2010)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017)
20. 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)
21. Liang, D., Zhou, X., Wang, X., Zhu, X., Xu, W., Zou, Z., Ye, X., Bai, X.: Point-mamba: A simple state space model for point cloud analysis. arXiv preprint [arXiv:2402.10739](https://arxiv.org/abs/2402.10739) (2024)
22. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model (2024), <https://arxiv.org/abs/2401.10166>
23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11976–11986 (2022)
24. Mante, V., Sussillo, D., Shenoy, K.V., Newsome, W.T.: Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature* **503**(7474), 78–84 (2013)
25. Miao, Z., Zhao, M., Zhang, X., Ming, D.: Lmda-net: A lightweight multi-dimensional attention network for general eeg-based brain-computer interfaces and interpretability. *Neuroimage* **276**, 120209 (2023)
26. Newson, J.J., Thiagarajan, T.C.: Eeg frequency bands in psychiatric disorders: a review of resting state studies. *Front. Hum. Neurosci.* **12**, 521 (2019)
27. Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., Ré, C.: S4nd: Modeling images and videos as multidimensional signals with state spaces. *Adv. Neural. Inf. Process. Syst.* **35**, 2846–2861 (2022)
28. Nidamanuri, J., Nibhanupudi, C., Assfalg, R., Venkataraman, H.: A progressive review: Emerging technologies for adas driven solutions. *IEEE Transactions on Intelligent Vehicles* **7**(2), 326–341 (2021)

29. Panwar, N., Pandey, V., Roy, P.P.: Eeg-cognet: A deep learning framework for cognitive state assessment using eeg brain connectivity. *Biomed. Signal Process. Control* **98**, 106770 (2024)
30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
31. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint [arXiv:2402.02491](https://arxiv.org/abs/2402.02491)* (2024)
32. Sheykhivand, S., Rezaii, T.Y., Meshgini, S., Makoui, S., Farzamia, A.: Developing a deep neural network for driver fatigue detection using eeg signals based on compressed sensing. *Sustainability* **14**(5), 2941 (2022)
33. Siddhad, G., Gupta, A., Dogra, D.P., Roy, P.P.: Efficacy of transformer networks for classification of eeg data. *Biomed. Signal Process. Control* **87**, 105488 (2024)
34. Siddhad, G., Iwamura, M., Roy, P.P.: Enhancing eeg signal-based emotion recognition with synthetic data: Diffusion model approach. *arXiv preprint [arXiv:2401.16878](https://arxiv.org/abs/2401.16878)* (2024)
35. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. *arXiv preprint [arXiv:2208.04933](https://arxiv.org/abs/2208.04933)* (2022)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
37. Wang, F., Wan, Y., Li, M., Huang, H., Li, L., Hou, X., Pan, J., Wen, Z., Li, J.: Recent advances in fatigue detection algorithm based on eeg. *Intelligent Automation & Soft Computing* **35**(3) (2023)
38. Wang, H., Zhu, X., Chen, P., Yang, Y., Ma, C., Gao, Z.: A gradient-based automatic optimization cnn framework for eeg state recognition. *J. Neural Eng.* **19**(1), 016009 (2022)
39. Wu, W., Gao, Y., Bienenstock, E., Donoghue, J.P., Black, M.J.: Bayesian population decoding of motor cortical activity using a kalman filter. *Neural Comput.* **18**(1), 80–118 (2006)
40. Yue, Y., Li, Z.: Medmamba: Vision mamba for medical image classification. *arXiv preprint [arXiv:2403.03849](https://arxiv.org/abs/2403.03849)* (2024)
41. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6848–6856 (2018)
42. Zheng, W.L., Lu, B.L.: A multimodal approach to estimating vigilance using eeg and forehead eeg. *J. Neural Eng.* **14**(2), 026017 (2017)
43. Zheng, Z., Wu, C.: U-shaped vision mamba for single image dehazing. *arXiv preprint [arXiv:2402.04139](https://arxiv.org/abs/2402.04139)* (2024)
44. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint [arXiv:2401.09417](https://arxiv.org/abs/2401.09417)* (2024)



EEG-Based Reaction Time Prediction Using Covariance Augmented 2D Convolutional Neural Network

Adarsh V. Parekkattil[✉], Sanjeev Kumar Varun[✉],
and Tharun Kumar Reddy Bollu^(✉)[✉]

Electronics and Communication Engineering, Indian Institute of Technology Roorkee,
Roorkee 247667, Uttarakhand, India
tharun.reddy@ece.iitr.ac.in

Abstract. Drowsy driving emerges as a major factor contributing to traffic accidents. Drowsiness is characterized by a feeling of tiredness and a compelling desire to sleep. It is evident through a gradual decrease in Reaction Time (RT) of the driver. Reaction Time (RT) refers to the duration taken for a person or system to respond to a given sudden unexpected stimuli or event. Accurate prediction of Reaction Time (RT) to unexpected events is crucial for enhancing safety and performance. Electroencephalogram (EEG), capturing the brain's electrical signals, demonstrates the most substantial correlation with drowsiness. Consequently, EEG is broadly recognized as a trustworthy tool for assessing drowsiness, fatigue, and overall performance. While many studies have utilized traditional machine learning and deep learning techniques to detect drowsiness or alertness from EEG data, there has been limited research focused on accurately predicting Reaction Time (RT). In this study, we aim to forecast drivers' reaction times using EEG data through a novel Covariance 2D CNN-LSTM framework. Here, the objective is to forecast Reaction Time (RT) based on a 5-second EEG trial that precedes it. The superiority of the proposed method was validated through regression metrics, specifically Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and was compared against current state-of-the-art methods.

Keywords: EEG · drowsiness detection · regression · CNN · LSTM · Reaction Time (RT) · RMSE · MAE

1 Introduction

Driving fatigue is a significant public safety issue, contributing to 15% to 20% of fatal traffic accidents. Efforts have been made to develop methods to detect fatigue and reduce related fatalities and economic losses. Drowsiness, characterized by fatigue and a strong urge to sleep, increases RT, the duration it takes to respond to a sudden stimulus. While many studies detect "drowsy" or "alert" states from EEG data, few focus on accurately predicting RT from EEG. This

study uses EEG data to directly predict RT, developing an advanced RT prediction model for drowsiness research.

Our study is organized as follows: Section 2 reviews recent works. Section 3 mentions about methods and materials, including EEG data collection, pre-processing, Covariance Matrix Transformation, proposed model architecture, and experimental setup. Sections 4 and 5 provide the results and discussions part, while the conclusion is covered in Section 6.

2 Related Work

Recent methods for detecting Reaction Time (RT) and fatigue have leveraged the extraction of various features from Electromyography (EMG) [7], Electrocardiogram (ECG) [4], Electrooculogram (EOG) [6] and Electroencephalogram (EEG) [8–10]. Melnicuk et al. recently provided a comprehensive review on driver state monitoring technologies that incorporate multiple features [5], for optimizing driving performance. Among the various features studied, physiological signals, especially EEG directly correlation with Reaction Time (RT) without being affected by external factors. EEG signals, in particular, have consistently been shown to be reliable biomarkers for detecting driving fatigue [5, 11].

Most studies on RT detection using EEG have focused on feature extraction within-subject analysis due to significant individual variations in behavioral performance and brain activity, which reduces the practicality of physiological-signal-based methods [2, 5]. Efforts to improve model transferability across subjects have assumed the same data distribution and feature space [12]. However, these feature extraction techniques often involve complex preprocessing algorithms with high computational demands and may fail with large datasets. Creating specifically tuned feature extractions for numerous subjects is impractical and these techniques often struggle to capture long-duration temporal dependencies, complicating transferability. An alternative is the CNN-based approach, which can generalize and extract features from large datasets with transferability [12]. By using heatmaps to capture temporal dependability, we can effectively convert this into an image regression problem.

In this study, we propose a unified Covariance-CNN-LSTM framework to predict RT from EEG signals, leveraging CNN’s superior automatic feature extraction capabilities with large datasets. Recognizing that EEG signals are temporal sequences with correlated consecutive moments, we address the limitation of traditional CNN models, which lack mechanisms to process sequential input correlations, leading to information loss. By introducing a Covariance matrix with CNN, we combine CNN with Covariance to extract fine-grained channel-wise temporal inter-dependencies effectively, a technique widely utilized in natural language processing. Our model recognizes that channel signals do not equally contribute to prediction and highlights the significance of correlations among multiple channel signals in detecting fatigue. This approach aims to create a practical in-vehicle system for identifying driving fatigue detection using RT, enhancing generalizability and efficiency in processing EEG data.

3 Methods and Materials

3.1 Lane Keeping Task (LKT) Dataset

The study uses the open-source Lane Keeping Task dataset by Cao et al. [1]. The experimental configuration includes driving simulations executed with virtual reality (VR) technology on an advanced driving simulator. The VR driving scenario replicated nighttime driving at 120 km/h on a straight, empty highway featuring two lanes in each direction. Random disruptions, termed as deviation onsets, were generated by the computer program, causing the vehicle to steer towards either side of the lane with equal probability.

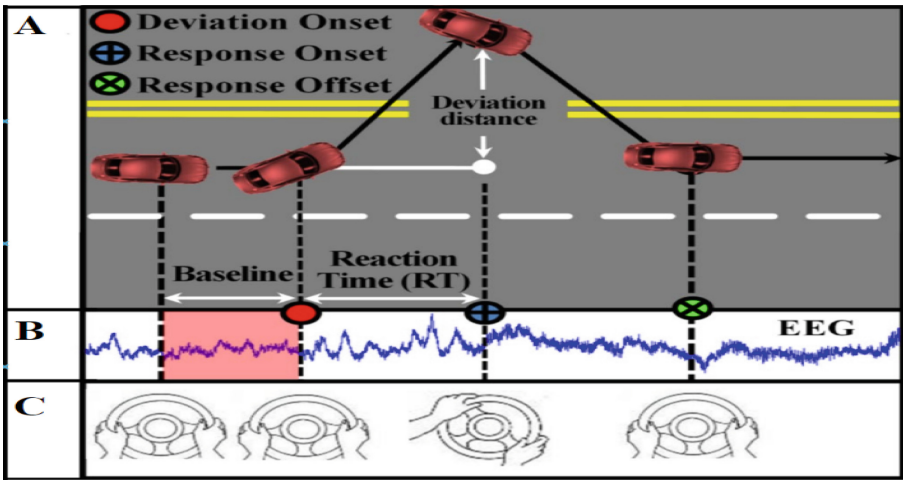


Fig. 1. (A) Paradigm with lane deviations caused by sudden stimuli. (B, C) EEG and behavioral data were collected simultaneously. Each trial records deviation onset, response onset, and deviation offset. Reaction time (RT) is defined as the period between deviation onset and response onset.

Participants were instructed to quickly steer the car back to the center of the cruising lane using the steering wheel (response onset) after each instance of the car deviation (deviation onset), and to maintain control once the car returned to the approximate lane center (response offset). A lane departure trial consists of three main events: deviation onset, response onset, and response offset as depicted in Fig. 1. The next lane departure trial occurs randomly, about 5 to 10 seconds after the current trial’s response offset. Reaction time (RT) for each lane departure trial is defined as the time between deviation onset and response onset. If the subject does not respond within 2.5 (1.5) seconds, the vehicle will drift towards the left (right) roadside without crashing, continuing forward against the curb unless the subject completely stops responding. No intervention is made if the subject falls asleep and stops responding. After the lapse, participants

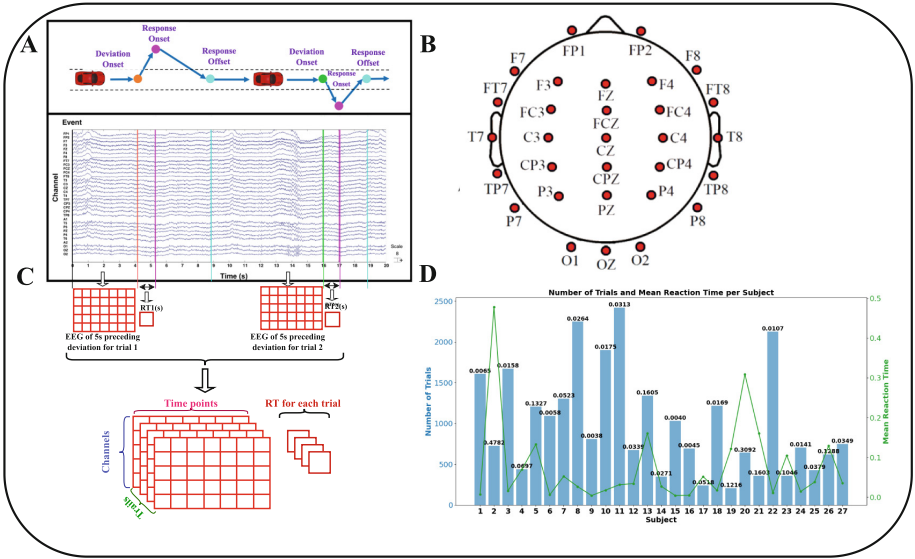


Fig. 2. (A) Behavioural performance and corresponding EEG signals with associated events. (B) Electrode locations for EEG recording. (C) EEG data format. (D) Subject-wise Trials and Mean RT plot

independently resume the task, steering the car back to the cruising position as soon as possible. The goal is to predict RT based on a 5-second EEG trial that precedes it. The study aimed to explore EEG patterns related to attention and performance changes during real-world drowsiness scenarios.

3.2 Data Collection and Preprocessing

The EEG data was gathered from 27 students at National Chiao Tung University (NCTU) in Taiwan, with an average age of 22.4 years and a standard deviation of 1.6 years. These experiments were sanctioned by the Institutional Review Board of Veterans General Hospital in Taipei, Taiwan. EEG signals were recorded using Ag/AgCl electrodes placed on a 32-channel Quik-Cap (Compumedical NeuroScan). Thirty electrodes were positioned following a modified international 10-20 system as depicted in Fig. 2 (B). The skin under the reference electrodes was prepared with Nuprep (Weaver and Co., USA) and cleaned with a 70% isopropyl alcohol swab before calibration. Electrode impedance was set below 5 k using NaCl-based conductive gel (Quik-Gel, Neuromedical Supplies). The EEG signals were amplified using the Scan NuAmps Express system (Compumedics Ltd., VIC, Australia) and recorded at a 500 Hz sampling rate with 16-bit quantization. Pre-processing involved bandpass filtering and artifact rejection, where raw EEG signals were filtered using 1 Hz high-pass and 50 Hz low-pass finite impulse response (FIR) filters. Eye blink artifacts were manually removed through visual inspection and further corrected using the Automatic

Artifact Removal (AAR) plug-in for EEGLAB. Behavioral performance and the corresponding EEG data are shown in Fig. 2 (A) and (C), with the average reaction time distribution of all 27 subjects across trials post-preprocessing shown in Fig. 2 (D) and Table 1.

Table 1. Subject-wise number of trials and average RT

Subject	1	2	3	4	5	6	7	8	9	10
Trials	1608	726	1673	438	1205	1094	1306	2248	809	1902
Mean RT	0.0065	0.4782	0.0158	0.0697	0.1327	0.0058	0.0523	0.0264	0.0038	0.0175
Subject	11	12	13	14	15	16	17	18	19	20
Trials	2422	675	1340	349	1033	695	239	1218	205	641
Mean RT	0.0313	0.0339	0.1605	0.0271	0.0040	0.0045	0.0518	0.0169	0.1216	0.3092
Subject	21	22	23	24	25	26	27			
Trials	360	2125	358	703	429	622	747			
Mean RT	0.1603	0.0107	0.1046	0.0141	0.0379	0.1288	0.0349			

3.3 Covariance Matrix Transformation

The pre-processed EEG time series data tensor is of shape $X \in \mathbb{R}^{D \times N \times K}$, where D represents the number of channels, N denotes the number of timepoints, and K indicates the number of trials, we convert each $D \times N$ matrix for a specific trial into a covariance matrix. For the k -th trial, the covariance matrix C_k is a $D \times D$ matrix, where each element $C_{ij}^{(k)}$ represents the covariance between the i -th and j -th channels:

$$C_{ij}^{(k)} = \frac{1}{N} \sum_{n=1}^N (X_{in} - \bar{X}_i)(X_{jn} - \bar{X}_j) \quad (1)$$

Here, \bar{X}_i and \bar{X}_j are the means of the i -th and j -th channels, respectively. The covariance matrix C_k thus encapsulates the pairwise covariances between all channels for the k -th trial. The range of values for elements in the covariance matrix depends on the scale of the original data. A positive value in $C_{ij}^{(k)}$ indicates a positive relationship between the i -th and j -th channels, suggesting that as the value of one channel increases, the other tends to increase as well. Conversely, a negative value implies an inverse relationship. The covariance matrix is symmetric ($C_{ij}^{(k)} = C_{ji}^{(k)}$), and the diagonal elements ($C_{ii}^{(k)}$) represent the variance of the i -th channel. We perform Covariance matrix conversion for all trials of a subject ie $1 \leq k \leq K$. A covariance matrix transformation converts pre-processed EEG time series data of a subject, $X \in \mathbb{R}^{D \times N \times K}$ to a collection of corresponding covariance tensor $C \in \mathbb{R}^{D \times D \times K}$, which can be represented as a collections heatmaps as shown in Fig. 3, i.e. no of trials = no of heatmaps = K . The covariance heatmaps of a subject are fed to the proposed 2D CNN-LSTM model to predict the reaction time (RT).

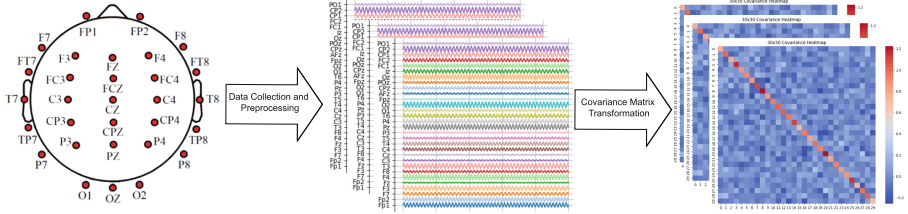


Fig. 3. Raw EEG collected using 10-20 electrode configuration with red markers indicating electrode placements. The recorded raw EEG is bandpass-filtered and transformed to covariance matrix heatmap, revealing inter-channel covariance relationships

3.4 Model Architecture

The proposed Covariance-CNN-LSTM model for RT prediction is shown in Fig. 4. It consists of both spatial and temporal feature extraction. The model begins with an input layer designed to handle reshaped covariance matrices of size 30x30x1, which are derived from the pre-processed time-series EEG signals. The first stage of the model involves three Conv2D layers, which serve as spatial feature extractors. The initial Conv2D layer uses 32 filters, each of size 3x3, resulting in an output dimension of 28x28x32 due to the reduction caused by the convolution operation (valid padding). This layer is followed by a ReLU activation function, mathematically represented as $Z^l = W^l * A^{l-1} + b^l$ and $A^l = \text{ReLU}(Z^l)$, where Z^l is the layer’s pre-activation output, W^l is the weight tensor, A^{l-1} is the activation from the previous layer, b^l is the bias vector, and ReLU is the activation function. Next, a MaxPooling2D layer with a pool size of 2x2 is applied, reducing the spatial dimensions to 14x14x32. MaxPooling layers serve to down-sample the input representation, reducing its dimensionality and allowing for the extraction of dominant features, while making the model more computationally efficient. The pooling operation is defined as $P_{i,j}^l = \max_{m,n \in R_{i,j}} A_{m,n}^l$, where $R_{i,j}$ is the receptive field at position (i, j) .

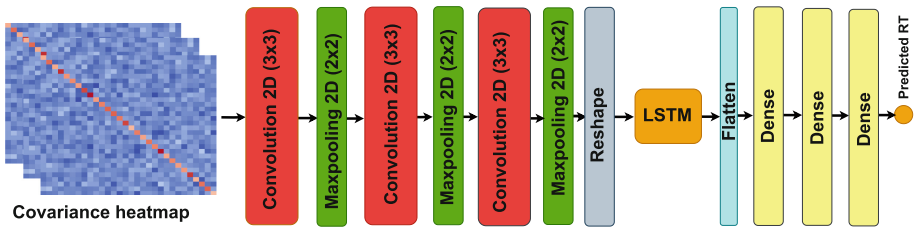


Fig. 4. Proposed Covariance 2D Convolutional Neural Network architecture for RT Prediction

The second Conv2D layer, with 64 filters of size 3x3, further processes these features, producing an output of 12x12x64, followed again by ReLU activation.

Another MaxPooling2D layer reduces this to 6x6x64. The third Conv2D layer employs 128 filters of size 3x3, generating a 4x4x128 output after convolution and ReLU activation. The subsequent MaxPooling2D layer reduces this to 2x2x128, capturing essential spatial features while significantly reducing dimensionality.

Table 2. Structure of the proposed architecture

Layers	Feature Map Size	Configuration	Output Shape
Input Layer	$30 \times 30 \times 1$	Input shape: $30 \times 30 \times 1$	(30, 30, 1)
Convolution Layer 1	$28 \times 28 \times 32$	3×3 conv, ReLU	(28, 28, 32)
MaxPooling Layer 1	$14 \times 14 \times 32$	2×2 max pooling	(14, 14, 32)
Convolution Layer 2	$12 \times 12 \times 64$	3×3 conv, ReLU	(12, 12, 64)
MaxPooling Layer 2	$6 \times 6 \times 64$	2×2 max pooling	(6, 6, 64)
Convolution Layer 3	$4 \times 4 \times 128$	3×3 conv, ReLU	(4, 4, 128)
MaxPooling Layer 3	$2 \times 2 \times 128$	2×2 max pooling	(2, 2, 128)
Reshape Layer	32×8	Reshape to 32×8	(32, 8)
LSTM Layer	32×64	LSTM with 64 units, return sequences	(32, 64)
Flatten Layer	2048	Flatten	(2048)
Dense Layer 1	128	Dense layer with 128 units, ReLU	(128)
Dense Layer 2	64	Dense layer with 64 units, ReLU	(64)
Output Layer	1	Dense layer with 1 unit, linear activation	(1)

The model then reshapes these feature maps into a 2D tensor of dimensions 32x8, preparing the data for temporal processing by the LSTM layer. The LSTM layer consists of 64 units and addresses the vanishing gradient problem common in traditional RNNs, capturing long-term dependencies in the data. The forget gate within the LSTM is defined as $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$, where f_t represents the forget gate activation, σ is the sigmoid function, W_f is the weight matrix, h_{t-1} is the previous hidden state, x_t is the current input, and b_f is the bias.

Following the LSTM layer, the model includes dense layers to perform nonlinear regression on the extracted spatio-temporal features. The first dense layer has 128 units, and the second dense layer has 64 units, both employing ReLU activation functions to introduce nonlinearity, represented as $h^l = \text{ReLU}(W^l h^{l-1} + b^l)$, enabling the model to approximate complex functions. Finally, the model concludes with a dense layer that outputs a single prediction value, making it suitable for regression tasks.

The overall configuration of the proposed architecture is shown in Table 2. ReLU activation ($f(x) = \max(0, x)$) is used throughout the network, except for the LSTM layer which uses tanh and sigmoid activations as per standard practice. ReLU helps mitigate the vanishing gradient problem and promotes sparsity in the activations. To prevent overfitting, L2 regularization is employed in the dense layers, defined as $L = L_0 + \lambda \sum \|W\|^2$, where L_0 is the original loss function, λ is the regularization strength, and $\|W\|^2$ is the L2 norm of the weight matrices.

Table 3. Computational Complexity of the Proposed Architecture

Layer Type	Output Shape	Equation for Mult-Adds	Mult-Adds	Parameters
Input	(30, 30, 1)	-	-	0
Conv2D (32x3x3)	(28, 28, 32)	$K \times F^2 \times C_{in} \times W \times H$	752,640	$K \times F^2 \times C_{in} + K = 320$
MaxPooling2D	(14, 14, 32)	-	-	0
Conv2D (64x3x3)	(12, 12, 64)	$K \times F^2 \times C_{in} \times W \times H$	1,327,104	$K \times F^2 \times C_{in} + K = 18,496$
MaxPooling2D	(6, 6, 64)	-	-	0
Conv2D (128x3x3)	(4, 4, 128)	$K \times F^2 \times C_{in} \times W \times H$	589,824	$K \times F^2 \times C_{in} + K = 73,856$
MaxPooling2D	(2, 2, 128)	-	-	0
Reshape	(32, 16)	-	-	0
LSTM (64 units)	(32, 64)	$N_{in} \times N_{out}$	20,480	$4N_h(N_h + N_i) = 20,736$
Flatten	(2048)	-	-	0
Dense (128 units)	(128)	$N_{in} \times N_{out}$	262,144	$N_{in} \times N_{out} + N_{out} = 262,272$
Dense (64 units)	(64)	$N_{in} \times N_{out}$	8,192	$N_{in} \times N_{out} + N_{out} = 8,256$
Dense (Output)	(1)	$N_{in} \times N_{out}$	64	$N_{in} \times N_{out} + N_{out} = 65$
Total	-	-	2,960,448	384,001

Computational Complexity and Model Architecture Computational complexity of an architecture determines its training and inference times. It is predominantly determined by the number of multiply-add (Mult-Adds) operations and the number of parameters in each layer. The Mult-Adds represent the operations required to compute the layer’s output from its input, while the parameters represent the weights and biases learned during training. For instance, in a Conv2D layer with K filters of size $F \times F$, applied to an input of size $W \times H \times C_{in}$ (where K is the number of filters, F is the filter size, W is the input width, H is the input height, and C_{in} is the number of input channels), the number of Mult-Adds can be calculated as $\text{Mult-Adds} = K \times F^2 \times C_{in} \times W \times H$. Similarly, for a dense (fully connected) layer, the Mult-Adds can be computed as $\text{Mult-Adds} = N_{in} \times N_{out}$ (where N_{in} is the number of input units and N_{out} is the number of output units). The total number of trainable parameters θ in the model is given by $|\theta| = \sum_{l=1}^L (K \times F^2 \times C_{in} \times C + C) + 4N_h(N_h + N_i) + \sum_{k=1}^K (N_{in} \times N_{out} + N_{out})$, where K is the number of filters, F is the filter size, C_{in} and C are the number of input and output channels in the Conv2D layers, N_h is the number of LSTM units, N_i is the input size to the LSTM, and N_{in} and N_{out} are the number of input and output units in dense layers. Table 3 illustrates the computational complexity of the proposed model. The number of Mult-Adds is a significant measure of computational complexity, particularly for layers such as Conv2D and Dense layers. From the complexity of this model, we understand that convolutional layers significantly contribute to the computational load due to their large number of Mult-Adds, while dense layers also add considerable complexity with their parameters. The balance of these elements determines the overall efficiency and capacity of the model to learn and generalize from data.

3.5 Experimental Setup

In this study, the experimental setup was configured on a workstation equipped with an Intel(R) Core(TM) i3-7020U CPU, a GeForce MX 110 GPU, and 8 GB of DDR4 RAM (Table 4).

Table 4. Workstation configuration.

Software or Hardware	Specification
CPU	Intel(R) Core(TM) i3-7020U
GPU	GeForce MX 110
RAM	DDR4 8 GB
Python	3.11.5
TensorFlow	2.16.2
Keras	3.4. 1

Table 5. Hyperparameter configuration.

Hyperparameters	Values
Activation function	ReLU (Conv2D and Dense layers), Linear (output layer), tanh and sigmoid(LSTM)
Optimizer	Adam
Learning rate	0.001
Loss function	Mean Squared Error
Metrics	Mean Absolute Error (MAE), Root Mean Square Error (RMSE)
Batch size	32
Epochs	100

Table 6. Performance Metrics

Metric	Equation
Mean Absolute Error (MAE)	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
Mean Squared Error (MSE)	$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

The software environment included Python 3.11.5, TensorFlow 2.16.2, and Keras 3.4.1, ensuring compatibility and performance for deep learning tasks. The model’s hyperparameters were carefully selected to optimize performance. Convolutional and dense layers utilized the ReLU activation function, while the output layer used a linear activation function (Table 5). The model is trained using the Adam optimizer with an initial learning rate of 0.001, and the loss function is mean squared error (MSE), given by $MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$, where y_i are the true values and \hat{y}_i are the predicted values. Training is conducted over 100

epochs with a batch size of 32, and early stopping based on validation loss is used to prevent overfitting. Evaluation metrics included mean absolute error (MAE) root mean squared error (RMSE) and Mean Squared Error (MSE) as shown in Table 6, ensuring a comprehensive assessment of the model’s performance.

4 Results

4.1 Single Trial RT Prediction

The model demonstrates superiority in predicting single trial Reaction Times (RT), as evidenced by its strong performance metrics across both training and validation datasets as shown in Table 7 and Table 8. Fig. 5 shows the Single Trial learning curve and RMSE after 100 epochs for subject 1 and for remaining subjects are given in supplementary material Section 1. For the training set, the model achieves remarkably low error rates, with an average MSE of 0.0017, MAE of 0.0231, and RMSE of 0.0342, indicating that the model accurately fits the training data with minimal deviation and effectively captures the underlying patterns of reaction times.

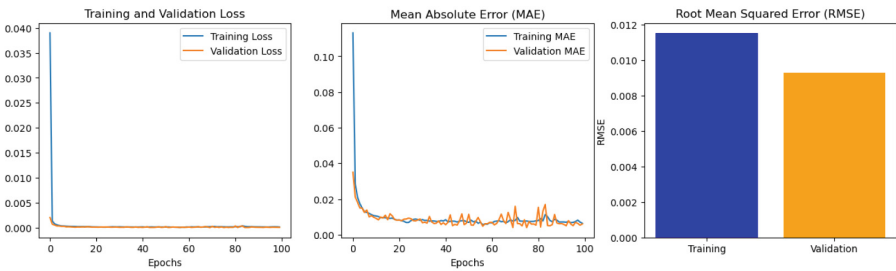


Fig. 5. Learning Curves and RMSE of Subject 1 for Single Trail RT prediction

Notably, subjects 2, 4, 14, and 22 show near-perfect training set performance, reflecting the model’s ability to learn and replicate individual-specific RT characteristics. On the validation set, the model maintains commendable performance, with an overall MSE of 0.00495, MAE of 0.0343, and RMSE of 0.0505, demonstrating its robustness and generalization capabilities, as it consistently delivers low error rates even on unseen data. Subjects such as 1, 4, 11, and 17 exhibit exceptionally low validation errors, highlighting the model’s precision in accurately predicting RT for different individuals. This consistency across both training and validation sets underscores the model’s reliability and effectiveness. Furthermore, the model’s performance metrics reveal its adaptability and sensitivity to diverse reaction time patterns among subjects, suggesting that it successfully captures the nuances and variations in reaction times. However, despite these strengths, there are some subjects, such as 3, 8, 10, 26, and 27,

Table 7. Subjectwise Performance Metrics for Single Trial Reaction Time Prediction

Subject	Training Set MSE	Training Set MAE	Training Set RMSE	Validation Set MSE	Validation Set MAE	Validation Set RMSE
1	0.0002	0.0071	0.0149	0.0001	0.0056	0.0089
2	0.0000	0.0026	0.0035	0.0009	0.0193	0.0296
3	0.0041	0.0512	0.0639	0.0169	0.1041	0.1300
4	0.0002	0.0071	0.0126	0.0000	0.0040	0.0046
5	0.0011	0.0249	0.0339	0.0063	0.0602	0.0795
6	0.0006	0.0159	0.0254	0.0016	0.0233	0.0405
7	0.0021	0.0236	0.0457	0.0051	0.0328	0.0712
8	0.0064	0.0596	0.0799	0.0331	0.1298	0.1819
9	0.0004	0.0124	0.0189	0.0028	0.0378	0.0526
10	0.0053	0.0548	0.0730	0.0283	0.1221	0.1682
11	0.0008	0.0139	0.0290	0.0002	0.0084	0.0141
12	0.0006	0.0192	0.0241	0.0025	0.0327	0.0501
13	0.0006	0.0192	0.0241	0.0025	0.0327	0.0501
14	0.0001	0.0044	0.0074	0.0004	0.0096	0.0193
15	0.0018	0.0242	0.0426	0.0011	0.0207	0.0330
16	0.0007	0.0162	0.0259	0.0041	0.0361	0.0643
17	0.0000	0.0027	0.0061	0.0000	0.0034	0.0045
18	0.0009	0.0181	0.0296	0.0005	0.0157	0.0222
19	0.0012	0.0213	0.0352	0.0023	0.0302	0.0474
20	0.0050	0.0529	0.0709	0.0083	0.0662	0.0910
21	0.0003	0.0103	0.0164	0.0019	0.0271	0.0431
22	0.0000	0.0021	0.0047	0.0003	0.0052	0.0160
23	0.0002	0.0126	0.0142	0.0003	0.0140	0.0167
24	0.0035	0.0358	0.0590	0.0087	0.0552	0.0934
25	0.0009	0.0128	0.0301	0.0017	0.0183	0.0416
26	0.0044	0.0500	0.0665	0.0458	0.1554	0.2140
27	0.0044	0.0500	0.0665	0.0458	0.1554	0.2140

Table 8. Overall Performance Metrics for Single Trial RT prediction for a subject after 100 epochs

Metric	Overall value
Mean Squared Error (MSE)	0.00495
Mean Absolute Error (MAE)	0.0343
Root Mean Squared Error (RMSE)	0.0505

who exhibit relatively higher validation errors, indicating that the model struggles to generalize well for these individuals. This variability could be attributed to individual differences in RT, data quality, or potential overfitting in certain cases. Overall, while the model demonstrates strong predictive capabilities and generalization performance, further refinement and investigation into subjects with higher error rates could enhance its robustness and accuracy.

4.2 Cross-Subject Performance

We evaluated Cross-Subject of the dataset Performance of a CNN-LSTM model for predicting RT using EEG data from multiple subjects. Each subject’s data was used as the validation set while the remaining data from other subjects formed the training set. This cross-validation approach ensured each subject’s

data was validated once, providing a comprehensive assessment of the model’s generalizability across different subjects. The performance after 100 epochs is detailed in Table 9 and Table 10. The learning curve for subject 1 is shown in Fig. 6 and for remaining subjects are given in supplementary material Section 2. The average performance metrics demonstrate the model’s strong prediction capabilities, with an average Cross-Subject Training RMSE of 0.1369 and a Validation RMSE of 0.0915. The average Cross-Subject Training MAE was 0.0673, while the Validation MAE was 0.0958, indicating the model’s ability to generalize well to new data with minimal error. Table 9 shows individual subject performance, where the Training RMSE ranged from 0.0516 to 0.0722 and the Validation RMSE from 0.0213 to 0.506. The Training MAE varied between 0.09 and 0.80, while the Validation MAE ranged from 0.04 to 0.159. These consistent results across subjects underscore the model’s robustness and reliability in predicting RTs from EEG data.

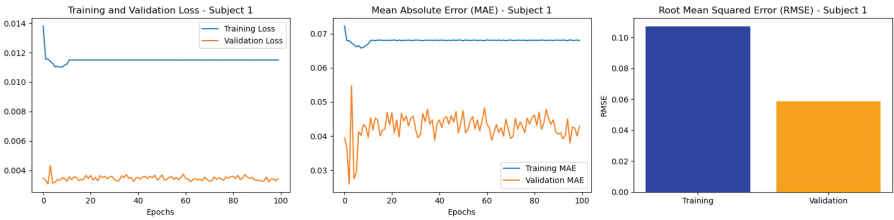


Fig. 6. Learning Curves and RMSE for Cross- Subject RT prediction of Subject 1

5 Discussions

The performance comparison between the proposed model and state-of-the-art models for single trial RT prediction is presented in Table 11. The results indicate that the proposed architecture significantly outperforms the existing models, including BP [10], ADG [3], and LGN [8], in terms of RMSE, MAE, and MAPE. Specifically, the proposed model achieves an RMSE of 0.0505, which is a notable improvement over the best-performing state-of-the-art model, LGN, which has an RMSE of 0.0751. Similarly, the MAE for the proposed model is 0.0343, markedly lower than the 0.097 achieved by the LGN model. Additionally, the proposed model’s MAPE of 3.43% demonstrates a substantial reduction in prediction error compared to the 9.7% MAPE of the LGN model.

Table 12 provides a comprehensive comparison of the proposed model’s performance for cross-subject reaction time prediction against other models. The proposed model demonstrates robust generalization capabilities, achieving an RMSE of 0.0915, an MAE of 0.0958, and a MSE of 0.0084. These results suggest that the proposed model maintains a high level of accuracy and consistency

Table 9. Cross-Subject Results for All Subjects after 100 Epochs

Subject	Training RMSE	Validation RMSE	Training MAE	Validation MAE
1	0.0683	0.0428	0.12	0.06
2	0.0668	0.0859	0.10	0.13
3	0.0689	0.0324	0.11	0.04
4	0.0516	0.4301	0.80	0.43
5	0.0702	0.0544	0.122	0.057
6	0.0715	0.0479	0.124	0.045
7	0.069	0.0213	0.11	0.039
8	0.0685	0.026	0.12	0.03
9	0.065	0.0776	0.114	0.1142
10	0.0698	0.0523	0.112	0.05
11	0.0686	0.0283	0.11	0.04
12	0.0693	0.506	0.121	0.05
13	0.069	0.051	0.113	0.044
14	0.071	0.0506	0.12	0.051
15	0.0722	0.0435	0.112	0.049
16	0.0635	0.1057	0.103	0.122
17	0.0683	0.0523	0.105	0.055
18	0.0675	0.059	0.106	0.058
19	0.0665	0.1248	0.104	0.159
20	0.0684	0.051	0.115	0.055
21	0.0673	0.0857	0.11	0.124
22	0.0689	0.0539	0.121	0.057
23	0.062	0.0749	0.102	0.142
24	0.060	0.2593	0.09	0.28
25	0.071	0.0535	0.12	0.05
26	0.0677	0.0607	0.109	0.08
27	0.0675	0.0547	0.102	0.06

Table 10. Average Performance Metrics for Cross- Subject Reaction Time(RT) prediction after 100 epochs

Metric	Training	Validation
MAE	0.0673	0.0958
RMSE	0.1369	0.0915
MSE	0.0187	0.0084

Table 11. Single Trial

Model	RMSE	MAE	MAPE
<i>BP</i> [10]	0.0853	0.148	14.8%
<i>ADG</i> [3]	0.0879	0.112	11.2%
<i>LGN</i> [9]	0.0751	0.097	9.7%
proposed	0.0505	0.0343	3.43%

Table 12. Cross-Subject

Model	RMSE
<i>EEGNet</i> [11]	0.029
<i>DeepCNN</i> [3]	0.028
<i>Shallow CNN</i> [11]	0.025
proposed	0.0915

across different datasets. Compared to rigorously tested and tuned EEGNet, DeepCNN, and Shallow CNN, which have lower RMSE values (0.029, 0.028, and 0.025 respectively), the proposed model shows high convergence with an RMSE of 0.0915 and excels in generalization across diverse data with faster inference times, essential for varied subject applications. While the RMSE is higher, the proposed model's performance in terms of MAE, MSE and convergence along with its ability to effectively handle raw EEG data with minimal preprocessing, makes it a highly valuable tool for accurate reaction time predictions in real-world scenarios with further parameter optimization.

6 Conclusion

The proposed Covariance Convolutional Neural Network framework excels in predicting reaction times, showing lower error rates and higher accuracy, especially in single trial predictions. Its impressive RMSE and MAE values highlight its precision and reliability. While it faces certain challenges in generalizing across diverse data, it still outperforms many existing models. Overall, the model is a significant advancement in reaction time prediction, with strong potential for practical applications and further development.




References

1. Cao et al., Z.: Multi-channel eeg recordings during a sustained-attention driving task. *Scientific data* **6**(1), 19 (2019)
2. Dehzangi, O., Taherisadr, M.: Driver distraction detection using mel cepstrum representation of galvanic skin responses and convolutional neural networks. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 1481–1486 (2018). <https://doi.org/10.1109/ICPR.2018.8545082>
3. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* **12**(7) (2011)
4. Kang, L., Kumar, J., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for document image classification. In: Proceedings of the 2014 22nd International Conference on Pattern Recognition. p. 3168–3172. ICPR '14, IEEE Computer Society, USA (2014)<https://doi.org/10.1109/ICPR.2014.546>, <https://doi.org/10.1109/ICPR.2014.546>

5. Melnicuk, V., Birrell, S., Crundall, E., Jennings, P.: Towards hybrid driver state monitoring: Review, future perspectives and the role of consumer electronics. In: 2016 IEEE Intelligent Vehicles Symposium (IV). pp. 1392–1397 (2016)<https://doi.org/10.1109/IVS.2016.7535572>
6. Murugan, S., Sivakumar, P.K., Kavitha, C., Harichandran, A., Lai, W.C.: An electro-oculogram (eog) sensor's ability to detect driver hypovigilance using machine learning. *Sensors* **23**(6) (2023).<https://doi.org/10.3390/s23062944>, <https://www.mdpi.com/1424-8220/23/6/2944>
7. Patricia, N., Tommasit, T., Caputo, B.: Multi-source adaptive learning for fast control of prosthetics hand. In: 2014 22nd International Conference on Pattern Recognition. pp. 2769–2774 (2014)<https://doi.org/10.1109/ICPR.2014.477>
8. Reddy, T.K., Arora, V., Behera, L., Wang, Y.k., Lin, C.T.: Multiclass fuzzy time-delay common spatio-spectral patterns with fuzzy information theoretic optimization for eeg-based regression problems in brain computer interface (bci). *IEEE Transactions on Fuzzy Systems* **27**(10), 1943–1951 (2019)<https://doi.org/10.1109/TFUZZ.2019.2892921>
9. Reddy, T.K., Arora, V., Gupta, V., Biswas, R., Behera, L.: Eeg-based drowsiness detection with fuzzy independent phase-locking value representations using lagrangian-based deep neural networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **52**(1), 101–111 (2022). <https://doi.org/10.1109/TSMC.2021.3113823>
10. Reddy, T.K., Arora, V., Kumar, S., Behera, L., Wang, Y.K., Lin, C.T.: Electroencephalogram based reaction time prediction with differential phase synchrony representations using co-operative multi-task deep neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence* **3**(5), 369–379 (2019). <https://doi.org/10.1109/TETCI.2018.2881229>
11. Reddy, T.K., Behera, L.: Driver drowsiness detection: An approach based on intelligent brain computer interfaces. *IEEE Systems, Man, and Cybernetics Magazine* **8**(1), 16–28 (2022). <https://doi.org/10.1109/MSMC.2021.3069145>
12. Weiss, K.R., Khoshgoftaar, T.M.: An investigation of transfer learning and traditional machine learning algorithms. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 283–290 (2016)<https://doi.org/10.1109/ICTAI.2016.0051>



Uncertainty-RIFA-Net: Uncertainty Aware Robust Information Fusion Attention Network for Brain Tumors Classification in MRI Images

Joy Dhar¹(✉) , Kapil Rana¹ , and Puneet Goyal^{1,2} 

¹ Indian Institute of Technology Ropar, Rupnagar, Punjab 140001, India
joy.22csz0003@iitrpr.ac.in

² NIMS Institute of Engineering and Technology, NIMS University, Jaipur 303121, Rajasthan, India

Abstract. Malignant brain tumors pose a significant global threat, emphasizing the critical need for efficient diagnostic methods utilizing MRI. Manual analysis of MRI images is labor-intensive and subjective, highlighting the necessity for faster and automated effective methods. In this paper, we propose an uncertainty-aware robust information fusion attention network model for precisely classifying brain tumors in MRI images. Our approach introduces a novel robust information fusion attention layer that learns enhanced representations by integrating global context with local information. We estimate the uncertainty in our model's predictions using the ensemble Monte Carlo dropout strategy. Our findings demonstrate outstanding performance, achieving accuracies of 98.37% on the Cheng dataset and 98.48% on the Nickparvar dataset in brain tumor MRI image classification tasks, while minimizing computational costs in terms of resource usage and inference time.

Keywords: Brain Tumors · MRI Images · Deep Learning · Attention Mechanism · Uncertainty Quantification

1 Introduction

The brain is a vital organ susceptible to life-threatening abnormalities such as tumors categorized as benign or malignant [6]. Benign tumors are treatable and non-fatal, whereas malignant tumors including gliomas and meningiomas, grow rapidly and rank among the top causes of global mortality [6], [2]. Treatment typically involves surgical intervention, radiotherapy and chemotherapy. However, diagnostic errors remain a significant contributor to mortality rates, necessitating the need for enhanced decision-support tools for medical specialists [6]. Magnetic resonance imaging (MRI) and computerized tomography (CT) scans are significant for accurate brain tumor detection. However, manual examination of these

These two authors contributed equally to this work.

scans is time-consuming, expensive and prone to human error, emphasizing the necessity for more efficient diagnostic methodologies [6].

Several deep learning (DL) approaches, including transfer learning (TL), lightweight convolutional neural networks (CNNs), and attention techniques, have achieved significant success in real-world image classification tasks. Considering medial image classification, deep learning based methods demonstrated significant performance across diverse medial image modalities including MRI images for the brain tumor classification [4]. These methods include transfer learning (TL) approaches to accelerate the training process [23], [22], Lightweight end-to-end CNN models for minimized computational resources and reduced inference time [12], [10], CNN with attention mechanisms for enhanced performance [4], [21], [9]. Most of these prior works utilized the Cheng dataset [7] and the Nickparvar dataset [13].

Most prior researchers often use TL approaches to enhance the performance of brain tumor classification using MRI images. For example, Deepak and Ameer [8] normalized image intensity and employed pre-trained GoogLeNet with modifications to classify brain tumour MRI images. In recent work, Zulfiqar et al. [23] fine-tuned the EfficientNetB2 model with data augmentation techniques. Zhu et al. [22] introduced evolutionary sparse DenseNets with evolution-based ensemble learning, which adapt towards higher density and efficiency across successive generations. Celik and Inik [6] developed a hybrid model based on TL-enabled CNN for feature extraction and Bayesian-optimized machine learning techniques for classification tasks.

Several researchers have introduced lightweight CNN models for brain tumor MRI image classification. For example, Hammad et al. [10] proposed a streamlined CNN architecture comprising three convolutional layers, max-pooling operations, and fully connected layers. Shahin et al. [16] devised a modified principal component analysis network, called MPCANet, which integrates unsupervised feature extraction with a supervised CNN for classification. Jaspin and Selvan [12] developed a multi-class CNN (MCCNN) model to improve classification performance.

Recent studies have explored the efficacy of attention networks in brain tumor MRI image classification. Xiao et al. [21] proposed ResNet34-CBAM, integrating the convolutional block attention module (CBAM) into a pre-trained ResNet framework, achieving a significant accuracy on small-scale MRI images. Billingsley et al. [4] proposed a lightweight CNN by designing a normalized attention mechanism to classify brain tumours with low inference time. Bodapati and Balaji [5] introduced Tumor Aware Net, an attention-based CNN trained end-to-end using sparse convolutional denoising autoencoder (SCDA) and neural-induced support vector classifier (NSVC). Most recently, Dutta et al. [9] proposed ARM-Net, a lightweight global attention-guided residual multiscale CNN designed to learn class-specific features. Oksuz et al. [14] introduced an attention-guided CNN architecture that integrates three pre-trained lightweight encoders with effective data augmentation strategies. Alzahrani [3] developed ConvAttenMixer, a transformer model integrating external attention and self-attention

with convolution mixtures, achieving superior spatial and channel-wise feature capture.

Prior methods [6], [23], [22], [8] have achieved significant performance in brain tumor MRI classification. However, these methods often necessitated substantial computational resources due to high learnable parameters and large input sizes, resulting in prolonged inference times. In response, recent works [12], [10], [16] have focused on developing lightweight CNN models to mitigate computational costs and reduce inference time, albeit sacrificing significant classification accuracy. Another approach explored by researchers [4], [21], [9], [5], [14], [3] involves integrating attention mechanisms into CNNs to enhance accuracy. However, they primarily emphasize minimizing computational resources and inference time, potentially compromising accuracy compared to traditional transfer learning approaches [23], [22]. This intuition emphasizes the need for a balanced approach to achieve high accuracy with minimal computational costs regarding resource usage and inference time, crucial for impactful healthcare research in brain tumor diagnosis. Motivated by these considerations, we propose an innovative uncertainty aware robust information fusion attention network (Uncertainty-RIFA-Net) model as a balanced approach to achieve excellent performance while obtaining efficient computational resources and reducing inference time for brain tumour MRI classification tasks. The main contributions of this work are as follows:

- We propose an innovative Uncertainty-RIFA-Net model for accurate brain tumour classification tasks in MRI images. We introduce a novel robust information fusion attention mechanism in our proposed network model.
- We estimate the uncertainty in our proposed model using an ensemble Monte Carlo dropout (EMCD) strategy.
- We conduct extensive experiments on the Cheng dataset [7] and the Nickparvar dataset [13] for brain tumor classification tasks using MRI images and compare the performance of our suggested approach with the prior state-of-the-art DL models.

2 Methodology

We introduce Uncertainty-RIFA-Net, a novel model designed for brain tumor classification in MRI images. This approach consists of three core phases: feature extraction, classification, and uncertainty quantification (UQ), as illustrated in Fig. 1(a). These phases synergistically improve performance by learning enhanced representations and estimating uncertainties in the prediction of model. In particular, the feature extraction phase is designed to capture enhanced global-local representations from the input features, thereby improving performance in brain tumor classification tasks. The classification phase then accurately categorizes brain tumors in MRI images. For uncertainty quantification, we employ an EMCD strategy to estimate prediction uncertainties in proposed RIFA-Net model.

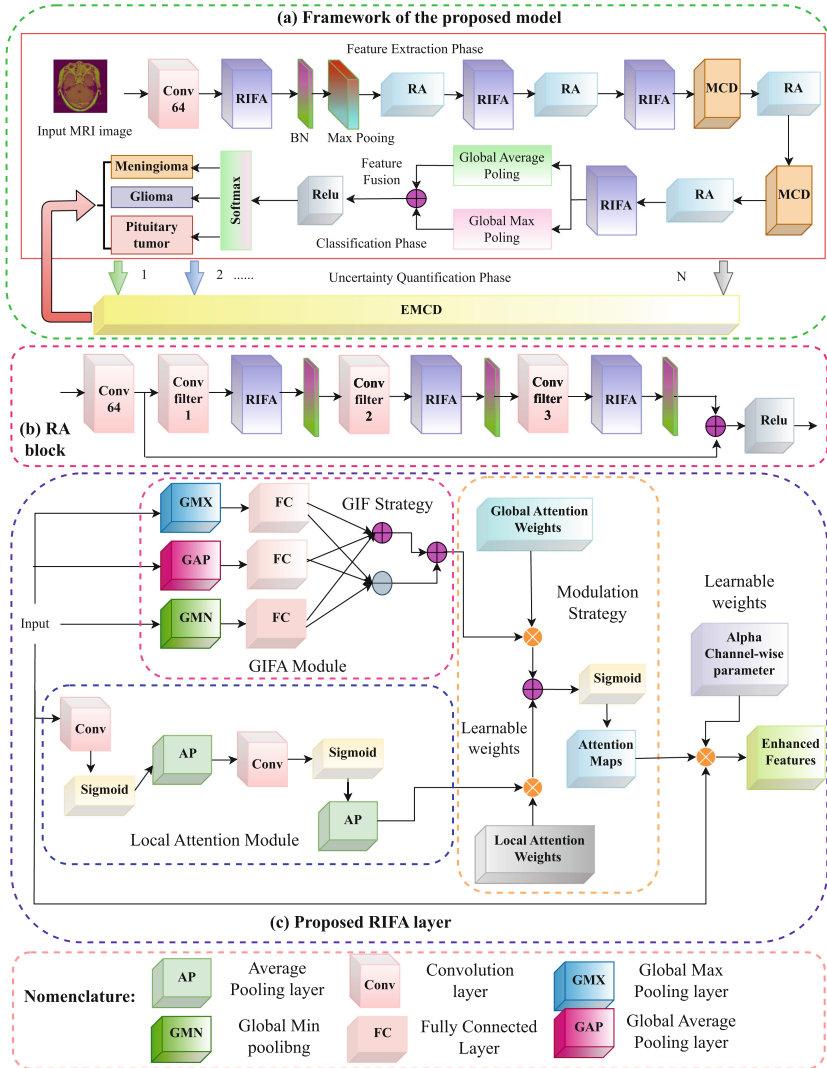


Fig. 1. (a) The proposed uncertainty aware robust information fusion attention network (Uncertainty-RIFA-Net) model consists of three phases: feature extraction, classification, and uncertainty quantification (UQ). The feature extraction phase captures both global contexts and fine-grained details, the classification phase performs brain tumor classification, and the UQ phase estimates prediction uncertainties. (b) The feature extraction phase includes a residual attention (RA) block, comprising convolutional layers, batch normalization layers, and the proposed robust information fusion attention (RIFA) approach. (c) The RIFA method comprises a global information fusion attention (GIFA) module, a local attention (LA) module, and a modulation strategy to enhance representation learning.

2.1 Feature Extraction Phase

In devising the feature extraction phase, we design four residual attention (RA) blocks to extract enhanced representations F' from the input features $f_i \in F$. Each RA block comprises four convolutional layers and three novel robust information fusion attention (RIFA) layers, placed after each convolutional layer (except the first one), followed by batch normalization layers, as shown in Fig. 1(b). This approach captures both global context and fine-grained details from the input features. Additionally, we employ an extra RIFA layer after each RA block (except the third) to learn enhanced representations from the given input.

RIFA Method In this study, we introduce the RIFA approach within and after each RA block (excluding the third) of the feature extraction phase in our base CNN model. This method effectively learns enhanced representations F' from input data, thereby improving the performance of our proposed model in brain tumor classification tasks, as depicted in Fig. 1(c). The pseudo-code of our proposed RIFA method is detailed in Algorithm 1.

The motivation for developing the RIFA method draws inspiration from the CBAM [19] and the channel spatial attention module (CSAM) [20]. CBAM focuses on learning spatial and channel information to enhance representation power, while CSAM captures global and local information from input data to improve learning across global and local contexts. However, our proposed RIFA method diverges from existing CBAM and CSAM approaches in three key aspects. Firstly, we introduce a global information fusion attention (GIFA) module that consists of global minimum γ_{mn} , global maximum γ_{mx} , and global average γ_{avg} pooling layers applied to the input features x . This contrasts with CSAM and CBAM, which use γ_{mx} and γ_{avg} layers for learning global information and β_{mx} and β_{avg} layers for channel information learning, respectively. Specifically, our GIFA module leverages all three pooling layers to learn diverse global information: minimum, maximum, and average. The resulting diverse information is fused using a global information fusion (GIF) strategy to enhance global information learning.

Secondly, we devise a local attention (LA) module that differs from the CBAM [19] and CSAM [20] approaches. The LA module employs a 1×1 convolution layer followed by average pooling to capture local information from input features, further enhancing the model's performance. Unlike CBAM and CSAM, LA module does not fuse outputs learned from each specified layer.

Thirdly, we introduce a modulation strategy that distinguishes itself from the CBAM [19] and CSAM [20] approaches, which do not include a modulation strategy similar to ours. Our approach integrates learnable global and local attention weights, ω_G and ω_L , alongside a fusion layer and sigmoid activation.

The primary intuition behind integrating a global minimum pooling layer and a global information fusion strategy within the GIFA module is to mitigate information loss by learning enhanced global information, thereby improving the performance of the model. Conversely, the main intuition behind using learnable attention weights, followed by a fusion layer and sigmoid attention within the

modulation strategy, is to learn robust global-local attention maps. These maps highlight important features and suppress less relevant ones, thereby improving overall performance.

Given input feature maps $F \in \mathbb{R}^{H \times W \times C}$, where H , W , and C denote the height, width, and number of channels respectively, the RIFA approach aims to enhance representation learning as F' . This method involves element-wise multiplication of global-local attention maps A with F , alongside channel-wise learnable parameters α_C , which adjust the importance of each channel during training, as shown in Equation 1.

$$F' = F \times A \times \alpha_C \quad (1)$$

To learn global-local attention feature maps A , we first develop two attention modules: the global information fusion attention module ρ and the local attention module ϑ . Second, we employ a modulation strategy. These components enhance the model's representation power by selectively highlighting important global and local features in the input data while suppressing less important ones.

Global Information Fusion Attention Module The global information fusion attention module comprises three global pooling layers: γ_{mn} , γ_{mx} , and γ_{avg} , which learn diverse global information such as minimum, maximum, and average. Each pooling layer is followed by a fully connected layer, η , to compress the learned global information. The GIF strategy then fuses the resultant global information to learn enhanced global information, φ . Specifically, the GIF strategy performs parallel element-wise operations, including addition and subtraction, on the variant global information to learn both the added global information, ζ , and the subtracted global information, λ , as shown in Equations 2-3. The resultant information is further fused to achieve enhanced global information, φ , as exhibited in Equation 4.

$$\zeta = \sum_{i=1}^3 \eta_i(\gamma_{\text{op}_i}(F)), \quad \text{where } \gamma_{\text{op}_i} \in \{\gamma_{mx}, \gamma_{avg}, \gamma_{mn}\} \quad (2)$$

$$\lambda = \eta_1(\gamma_{mx}(F)) - \eta_2(\gamma_{avg}(F)) - \eta_3(\gamma_{mn}(F)), \quad \text{where } \eta_1, \eta_2, \eta_3 \in \eta \quad (3)$$

$$\varphi = \rho(F) = (\zeta \oplus \lambda), \quad \oplus \in \text{fusion layer by addition} \quad (4)$$

Local Attention To develop the local attention module ϑ , we use a convolution layer, α , with sigmoid activation, σ , followed by an average pooling layer to compress the learned local information, ψ . We repeatedly use these layers r times to preserve essential features and thereby learn fine-grained details from input features, F , such that $0 \leq r \leq 1$, as exhibited in Equation 5.

$$\psi = \vartheta(F) = \left[\frac{1}{H \times W} \sum_{i,j} (\sigma(\alpha(F))) \right]_r \quad (5)$$

Algorithm 1: Robust Information Fusion Attention (RIFA) Method

- 1: **Input:** Input features, $f_i \in F$; where $F \in \mathbb{R}^{H \times W \times C}$
 - 2: **Output:** Enhanced representations, F'
 - 3: **Procedure:**
 - 4: /* To global-local attention feature maps A , */
 - 5: **Design global information fusion attention (GIFA) module, ρ :**
 - 6: /* Learn diverse global information from F using various pooling layers followed by fully connected layers */
 - 7: $Max = \eta_1(\gamma_{mx}(F))$
 - 8: $Avg = \eta_2(\gamma_{avg}(F))$
 - 9: $Min = \eta_3(\gamma_{mn}(F))$
 - 10: /* Use global information fusion (GIF) strategy to fuse all learned global information to capture enhanced global information, φ */
 - 11: $\zeta = Max + Avg + Min$
 - 12: $\lambda = Max - Avg - Min$
 - 13: $\varphi = \rho(F) = (\zeta + \lambda)$
 - 14: **Design local attention (LA) module, ϑ to learn local information:**
 - 15: $\psi = \vartheta(F) = \left[\frac{1}{H \times W} \sum_{i,j} (\sigma(\gamma(F))) \right]_r$
 - 16: **Design modulation strategy to highlight important features while suppressing less significant ones:**
 - 17: /* Use learnable weights ω_G and ω_L to enhance global and local information */
 - 18: $M_G = \varphi \times \omega_G$
 - 19: $M_L = \psi \times \omega_L$
 - 20: /* Fuse global and local information followed by sigmoid activation */
 - 21: $A = \sigma(M_G + M_L)$
 - 22: /* To learn enhanced representations */
 - 23: $F' = F \times A \times \alpha_C$
 - 24: **return F'**
 - 25: // return
-

Modulation In the modulation strategy, we use learnable weights ω_G and ω_L to adjust the importance of each global and local attention component during training. The global attention weight ω_G is initialized to one, while the local attention weights ω_L are initialized based on channel information. These weights perform element-wise multiplication with the outputs of ρ and ϑ , enhancing the learning of global (M_G) and local (M_L) information, respectively, as shown in Equation 6. The resulting global and local information is fused to strengthen the acquisition of robust global-local information. Additionally, a sigmoid activation function is applied to learn robust global-local attention maps A , ensuring attention scores lie within the range of 0 to 1, as exhibited in Equation 7. This effectively highlights important features while suppressing less significant ones, thereby adeptly capturing both long-range dependencies and fine-grained details.

$$M_G = \varphi \times \omega_G \text{ and } M_L = \psi \times \omega_L \quad (6)$$

$$A = \sigma(M_G \oplus M_L) \quad (7)$$

This study activates the learnable attention weights in the RIFA layer using a boolean layer parameter. These weights are optimized using a back-propagation strategy. Specifically, we compute the gradient of the loss δL with respect to the global and local attention weights along with channel-wise weights, specified as follows.

$$\frac{\delta L}{\delta \omega_G} = \frac{\delta L}{\delta A} \frac{\delta A}{\delta \omega_G} \text{ and } \frac{\delta L}{\delta \omega_L} = \frac{\delta L}{\delta A} \frac{\delta A}{\delta \omega_L} \quad (8)$$

$$\left. \begin{array}{l} \frac{\delta A}{\delta \omega_G} \\ \frac{\delta A}{\delta \omega_L} \end{array} \right\} = A \times (1 - A) \text{ and } \frac{\delta L}{\delta A} = \frac{\delta L}{\delta M_G} + \frac{\delta L}{\delta M_L} \quad (9)$$

where $\delta \omega_G$, $\delta \omega_L$, δA , δM_G , and δM_L are back-propagations of the gradients. Now, we compute the gradient loss concerning learnable α_C parameter as follows.

$$\frac{\delta L}{\delta \alpha_C} = \beta \times \frac{\delta L}{\delta A} \times F \times A(1 - A), \quad (10)$$

where $\delta \alpha_C$ is the back-propagation of the gradient, and β is the constant in back-propagation.

Throughout the optimization process, our proposed approach iteratively updates the parameters. ω_G , ω_L , and α_C based on their computed gradients, aiming to minimize the overall loss of our proposed Uncertainty-RIFA-Net model. These learnable parameters undergo updates at step $t + 1$ for layer l as follows.

$$\left. \begin{array}{l} \omega_G^{t+1} = \omega_G^t - l_r \times \frac{\delta L^t}{\delta \omega_G}, \\ \omega_L^{t+1} = \omega_L^t - l_r \times \frac{\delta L^t}{\delta \omega_L}, \\ \alpha_C^{t+1} = \alpha_C^t - l_r \times \frac{\delta L^t}{\delta \alpha_C} \end{array} \right\} \quad (11)$$

The resulting enhanced representations are then passed to the subsequent layers to perform operations, as illustrated in Fig. 1(a).

2.2 Classification Phase

In the classification phase, we employ a γ_{mx} and γ_{avg} layers, along with a feature fusion layer to fuse their output features. Subsequently, a SoftMax classifier layer is used to classify brain tumor in MRI images.

2.3 Uncertainty Quantification Phase

After developing the RIFA-Net model, we perform uncertainty quantification using an EMCD strategy. This involves generating N models and utilizing k input samples to estimate prediction uncertainties y'' . This strategy significantly enhances the reliability of our model’s predictions. Monte Carlo dropout layers are applied after the second and third RA blocks, followed by the RIFA layer and subsequent RA block, as shown in Fig. 1(a).

Initially, in EMCD, we develop the RIFA-Net model with MCD, denoted as $m(\cdot)$, and run each model N times. We use model ensembling to acquire predictions from N trained models. These models have various weight distributions and initialized weights, which significantly enhances model performance. After model training, we utilize the Monte Carlo equation with $K = 30$, as shown in Equation 13. This helps obtain our model predictions through different stochastic paths using MCD, d_u , creating randomness in our model architecture.

We aggregate all predictions by calculating the mean for each sample to obtain the softmax probabilities, y_u . This creates an ensemble of diverse models to enhance overall performance. During testing, we run the RIFA-Net model 30 times per sample. The average prediction is taken as the final prediction, thereby estimating prediction uncertainty in our model.

$$y_u = \text{softmax}(m(F, d_u)) \quad \text{and} \quad y' = \frac{1}{N} \sum_{u=0}^{N-1} y_u \quad (12)$$

$$y'' = \text{argmax}(y') \quad (13)$$

3 Experiments and Results

3.1 Experimental Setup

This section elaborates on the employed datasets, comparison models, and implementation details.

Dataset Experiments were conducted on two publicly available MRI medical imaging datasets for brain tumor classification tasks: the Cheng dataset [7] (D1) and the Nickparvar MRI dataset [13] (D2). Additionally, to evaluate the potential effectiveness of our RIFA-Net approach in a different modality, specifically CT-scan, we employed the brain stroke CT image dataset [1] (D3) for brain stroke classification in CT images.

For the MRI modality, the Cheng dataset [7] consists of 3064 T1-weighted MRI images collected from 233 patients, categorized into Meningioma (labeled as 0), Glioma (labeled as 1), and Pituitary (labeled as 2) tumor classes. Originally in .mat format, this dataset was converted to .png. The Nickparvar dataset [13] combines the Figshare, SARTAJ, and BrH35 datasets, containing 7023 MRI images. These include Glioma (324 test and 1297 training images, labeled as 0),

Meningioma (329 test and 1316 training images, labeled as 1), No Tumor (400 test and 1600 training images, labeled as 2), and Pituitary (351 test and 1406 training images, labeled as 3).

For the CT-scan modality, the brain stroke CT image dataset [1] comprises 2501 images, including 1551 normal cases (labeled as 0) and 950 stroke cases (labeled as 1). These images have an in-plane resolution of 0.5 mm and a slice thickness ranging between 4 and 5 mm, covering the entire brain. They were collected from over 1000 patients in a clinical setting.

All MRI and CT-scan images in these datasets were resized to $128 \times 128 \times 3$. Datasets D1 and D3 used 80% of the images for training and 20% for testing, while D2 was by default pre-split into training and testing sets.

Comparison Models and Implementations Details In this study, we benchmarked our RIFA-Net model against traditional CNNs, baseline CNNs, and state-of-the-art methods using datasets D1-D3 [1, 7, 13] for classifying brain tumors and strokes in MRI and CT-scan modalities. The state-of-the-art methods included in [3–6, 8–10, 12, 14–16, 18, 21]. Additionally, we applied the EMCD method to quantify uncertainty in our RIFA-Net model predictions.

In this study, all implementations were carried out on an NVIDIA GeForce RTX 2080 Ti coprocessor. Models were trained for 200 epochs using cross-entropy loss with a batch size of 64. The Adam optimizer was utilized with an initial learning rate set to 0.001. We employed a Reduce Learning Rate on Plateau scheduler, which reduces the rate by a factor of 0.2 when performance plateaus, with a patience parameter of 5 and a lower bound for the learning rate set at $2e-4$. For uncertainty quantification, we employed the EMCD module to generate $n = 5$ models. In the EMCD module, dropout rates of 0.5 were applied for the first Monte Carlo dropout layer and 0.25 for subsequent layers, as illustrated in Fig. 1(a). Each RGLA layer consisted of three convolutional layers: 64, 64, and 256 for layers 1 and 2; and 128, 128, and 512 for the remaining RGLA layers.

3.2 Performance Comparison Results

In this section, we evaluate RIFA-Net’s performance on D1 and D2 datasets of the MRI modality for brain tumor classification tasks, as presented in Tables 1 and 2. We compare RIFA-Net’s performance with state-of-the-art methods, including TL-CNN [8], L2-SA [4], Lightweight-CNN [10], MPCANet [16], MCCNN [12], TL-ResNet34-CBAM [21], Tumour-AwareNet [5], ARM-Net [9], AttentionGuided-CNN [14], ConvAttenMixer [3], EfficientNetB0-SVM, and CNN-KNN [6] on D1 and D2 datasets.

Results and Discussion: In Tables 1 and 2, the experimental results suggest that the RIFA-Net model achieved desirable performances of 98.37% and 98.54% on the D1 and D2 datasets, respectively. Compared to traditional CNNs such as VGG16 and ResNet18, and our baseline CNN models with or without CBAM, RIFA-Net showed significant performance enhancements ranging from 1.9% to 20.33% on the D1 and D2 datasets, respectively. Furthermore,

Table 1. Performance comparison with DL models and prior state-of-the-art methods to classify brain tumours using MRI images conducted on D1: Cheng dataset [7]. Here, **bold** values indicate the highest performance for this dataset.

Model	Year	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Inference Time (in seconds)	Parameters used (in millions)	Augmentation
ResNet50 [11],	2015	81.89	78.78	80.88	77.83	0.413	23.85	No
VGG16 [17]	2014	94.29	93.29	93.90	93.58	0.059	7.949	No
Deepak and Ameer (TL-CNN) [8]	2019	93.00	-	-	-	-	-	-
Bodapati and Balaji (Tumour-AwareNet)[5]	2023	96.25	95.65	95.96	95.78	-	2.40	No
Jaspin and Selvan (MCCNN) [12]	2023	96.40	-	95.90	-	-	13.10	-
Billingsley et al. (L2-SA) [4]	2023	96.57	-	-	-	0.0019	7.293	No
Dutta et al. (ARM-Net) [9]	2023	96.64	96.40	96.09	96.20	-	1.13	Yes
Shahin et al. (MPCANet) [16]	2023	96.73	-	-	-	-	-	-
Hammad et al. (Lightweight-CNN) [10]	2023	96.86	97.00	97.00	97.00	0.02	2.45	-
Oksuz et al. (AttentionGuided-CNN) [14]	2023	97.02	94.95	-	94.91	-	6.63	Yes
Xiao et al. (TL-ResNet34-CBAM) [21]	2021	97.44	98.25	96.87	97.33	-	-	-
Our work (Base CNN)	-	93.80	92.84	93.70	93.22	0.060	3.806	No
Our work (Base CNN-CBAM)	-	95.10	94.04	95.21	94.56	0.094	4.183	No
Our Work (Proposed Uncertainty-RIFA-Net)	-	98.37	97.82	98.54	98.16	0.07	7.22	No

our proposed model outperformed existing state-of-the-art methods in terms of accuracy (0.54% to 5.37%), precision (0.38% to 2.87%), recall (0.35% to 3.08%), and F1-score (0.37% to 2.35%) on the D1 and D2 datasets [7, 13].

Table 2. Performance comparison with DL models and prior state-of-the-art methods to classify brain tumours using MRI images conducted on D2: Nickparvar MRI dataset [13]. Here, **bold** values indicate the highest performance for this dataset.

Model	Year	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Inference Time (in seconds)	Parameters used (in millions)	Augmentation
ResNet50 [11]	2015	89.93	89.5	89.25	89.26	0.035	23.85	No
VGG16 [17]	2014	96.34	96.21	96.03	96.09	0.034	14.78	No
Celik and Inik (CNN-KNN) [6]	2024	97.15	97.00	97.00	97.00	-	10.9	No
Celik and Inik (EfficientNetB0-SVM) [6]	2024	97.93	98.00	98.00	98.00	-	5.3	No
Alzahrani (ConvAttenMixer) [3]	2023	97.94	98.58	95.27	96.65	-	2.01	Yes
Our work (Base CNN)	-	96.34	96.21	96.03	96.08	0.0353	3.806	No
Our work (Base CNN-CBAM)	-	96.49	96.48	96.31	96.27	0.036	4.183	No
Our Work (Proposed Uncertainty-RIFA-Net)	-	98.48	98.38	98.35	98.37	0.038	7.22	No

The experimental results highlight the limitations of existing methods in achieving robust performance for brain tumor classification in MRI images. Specifically, methods such as [6, 8, 12, 21] often require high computational resources, leading to extended inference times but still exhibit limited performance due to facing information loss issues. Conversely, other approaches like [3–5, 9, 10, 14, 16] prioritize lightweight models at the expense of classification performance, as shown in Tables 1 and 2. These methods overlook the crucial need for high-performing models in the healthcare domain, focusing primarily on lightweight models. In contrast, our method represents a balanced approach, prioritizing both desirable performance and computational efficiency by minimizing computational costs, thereby reducing inference time for brain tumor classification.

Although RIFA-Net requires slightly more parameters and inference time compared to Base CNN and Base CNN-CBAM models, as shown in Tables 1 and 2. However, the RIFA-Net approach offers enhanced performance without

extensive computational overhead. This includes avoiding a higher number of parameters, augmentation strategies (except resizing), and larger input image sizes.

Table 3. Uncertainty quantification of our proposed model predictions with the Base-CNN-CBAM attention model on D1: Cheng dataset [7] and D2: Nickparvar MRI dataset [13]. Here, **bold** values indicate the highest performance for these datasets.

Model	Accuracy	Precision	Recall	F1	Dataset
Base CNN-CBAM	95.43	95.12	94.30	94.67	D1
Proposed Method	97.88	97.10	98.32	97.85	D1
Base CNN-CBAM	96.54	96.50	96.4	96.5	D2
Proposed Method	97.87	97.78	97.81	97.78	D2

3.3 Impact on Uncertainty Quantification

The experimental results discussed in Sub-Section 3.2 indicate that our RIFA-Net performs well, suggesting its potential use for the automatic classification of brain tumors in MRI images by clinical practitioners. We believe there is an urgent need for new, efficient, intelligent deep learning models for MRI data analysis in brain tumor classification, particularly those that include significant uncertainty estimation. To achieve this, we applied the uncertainty quantification method, known as EMCD, to quantify the uncertainty of our RIFA-Net model predictions.

Table 3 shows that our proposed model achieved significant performance and outperformed the Baseline-CNN-CBAM model when considering uncertainty for the D1 and D2 MRI datasets. Comparing our model with and without UQ, we found that our proposed model with UQ exhibited a slight performance drop compared to RIFA-Net without UQ. Conversely, the Baseline-CNN-CBAM model also showed a slight performance decline when UQ was applied. Our proposed model maintained stable performance, with minimal drops in accuracy (0.49% and 0.61%), precision (0.72% and 0.60%), recall (0.22% and 0.54%), and F1-score (0.58% and 0.59%) compared to RIFA-Net without UQ on the D1 and D2 datasets [7, 13].

3.4 Qualitative Analysis

We conducted a qualitative analysis using the Score-CAM technique to evaluate the effectiveness of the RIFA-Net model in classifying brain tumors for D1 and D2 datasets, as shown in Fig. 2. The Score-CAM technique visualized attention maps that highlighted the most important regions in MRI images contributing to our model’s decision for the target class. These maps give higher importance

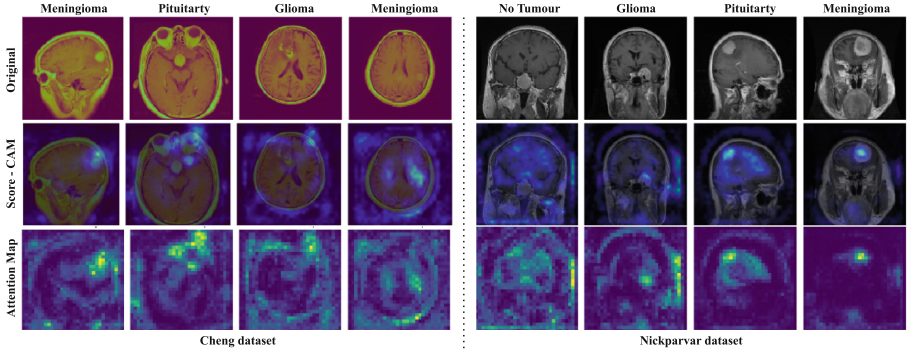


Fig. 2. Visual representation of our proposed method’s predictions, highlighting the most important regions in MRI images for classifying various types of brain tumors using the Score-CAM technique on the D1: Cheng dataset [7] and D2: Nickparvar dataset [13]. The top row shows the original images, the middle row shows the Score-CAM images, and the bottom row shows the attention map images.

to areas with higher prediction scores for the D1 and D2 datasets [7] and [13]. Fig. 2 illustrates the most important features used by our model to identify each data class separately for the D1 and D2 MRI datasets.

4 Ablation Study

The impact of individual components within the RIFA method was evaluated through six approaches for brain tumor classification in MRI images (Table 4). Each approach incorporates one or more key components: global attention (GA), global information fusion (GIF), and local attention (LA), with one approach serving as a baseline CNN without CBAM. A modulation strategy is integrated into all approaches. The six approaches are as follows: A1, the Base-CNN model; A2, incorporating global attention with global average pooling; A3, employing local attention only; A4, combining global and local attention without GIF; A5, combining global attention with GIF but without local attention; and A6, integrating all components to form the RIFA approach. The effectiveness of these components is demonstrated in Table 4.

Results and Discussion: The experimental results in Table 4 shows that our proposed method, RIFA (A6), which employs global attention with the GIF strategy and local attention, outperforms the baseline approaches (A1-A5). These baseline models lack components necessary for learning enhanced representations, leading to information loss and reduced performance. The GIF strategy in the GIFA module is crucial for improved performance, as it learns diverse global information from multiple pooling layers and fuses these representations to create enhanced global information. Local attention further captures fine-grained details, boosting model robustness. Consequently, RIFA significantly surpasses both baseline models and state-of-the-art methods in brain tumor classification.

Table 4. Experimental analysis of the individual components of our proposed robust information fusion attention (RIFA) method for enhanced representation learning. Table 4 presents experiments evaluating components of our proposed RIFA approach to learn enhanced representations and thereby significant performance improvement on the D2: Nickparvar dataset [13]. GA denotes global attention without global information fusion, GIF denotes global information fusion that fuse variant global information using parallel fusion (addition and subtraction) followed by addition, and LA represents local attention. Here, **bold** values indicate the highest performance for this dataset.

Approach	GA	GIF	LA	Method	Accuracy (%)	F1 score (%)
A1	No	No	No	RIFA	96.34	96.08
A2	Yes	No	No		96.70	96.34
A3	No	No	Yes		96.49	96.27
A4	Yes	No	Yes		98.09	97.96
A5	Yes	Yes	No		97.34	97.00
A6	Yes	Yes	Yes		98.48	98.37

5 Further Experiments of Brain Disease Analysis on CT-scan Modality

Primarily, this study presents brain tumor classification in MRI images using the RIFA-Net approach. However, we also demonstrate the effectiveness of RIFA-Net in classifying brain strokes using CT scans. This highlights its applicability across diverse medical imaging modalities for various brain diseases.

This section aims to highlight the effectiveness of our RIFA-Net model in the CT-scan modality for classifying brain strokes in CT-scan images, supported by basic experimental evidence presented in Table 5. Specifically, we utilized the D3 dataset [1] of the CT-scan modality to validate the desirable performance achieved by our RIFA-Net in brain stroke classification tasks. We also compared the performance of RIFA-Net with state-of-the-art methods such as OzNet-mrMR-NB [15] and Improved XGBoost [18] to demonstrate its significance among these existing approaches, as illustrated in Table 5.

Results and Discussion: In Table 5, the experimental results suggest that RIFA-Net achieved significant performance improvements with accuracy, precision, recall, and F1 score ranging from 99.92% to 99.94% on the D3 dataset. Compared to the aforementioned existing methods, our RIFA-Net model outperformed them by enhancing robustness by 1.5% to 2.94% on D3 dataset [1]. We applied the UQ strategy to quantify the uncertainty of our RIFA-Net model predictions on the D3 dataset, demonstrating the stability of our proposed learning model. RIFA-Net maintained stable performance, with minimal drops ranging from 0.41% to 0.49% on the D3 dataset [1].

Table 5. Performance comparison with existing methods, such as OzNetmRMRNB [15], Improved XGBoost [18], for brain stroke classification in CT-scan images using the D3: brain stroke CT image dataset [1]. **Bold** values indicate the highest performance for this dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Inference time (in ms)	Parameter used (in millions)
Improved XGBoost [18]	97.00	98.00	-	-	-	-
OzNetmRMRNB [15]	98.42	98.29	97.54	98.41	-	-
Uncertainty-RIFA-Net (Ours)	99.50	99.49	99.45	99.45	0.105	7.22
RIFA-Net (Ours)	99.92	99.94	99.90	99.92	0.095	7.22

Existing methods for acute brain stroke classification [15, 18] mainly use hybrid methods with CNNs local binary patterns Gabor filters and discrete wavelet transforms for feature extraction and employ traditional machine learning models like naïve Bayes (NB) and XGBoost for classification. However, these methods often lack robust performance due to limited focus on feature extraction and lack of enhanced representation learning with attention mechanisms. In contrast, our method in contrast emphasizes enhanced representation learning using the RIFA method for superior classification performance in acute brain stroke classification.

6 Conclusion

We propose the uncertainty-RIFA-Net model tailored for precise brain tumor classification in both MRI and CT-scan images. Our model incorporates a novel RIFA layer designed to enhance representations by capturing intricate long-range patterns and fine-grained details from input features. By leveraging EMCD, we quantify the uncertainty inherent in our model’s predictions. Our proposed model demonstrates significant accuracies of 98.37%, 98.48%, and 99.92% on the Cheng, Nickparvar, and CT image datasets, respectively. In comparative evaluations against baseline CNNs and state-of-the-art methods, our model consistently exhibits superior performance while maintaining low computational overhead. Future directions include extending our approach to diverse medical imaging modalities for more accurate disease classification, encompassing conditions like COVID-19, skin cancer, and diabetic retinopathy.

References


1. Brain stroke CT image dataset, <https://www.kaggle.com/datasets/afdirahman/brain-stroke-ct-imagedataset>
2. Brain Tumor - Statistics — cancer.net. <https://www.cancer.net/cancer-types/brain-tumor/statistics>, [Accessed 15-06-2024]
3. Alzahrani, S.M.: ConvAttenMixer: Brain tumor detection and type classification using convolutional mixer with external and self-attention mechanisms. *Journal of King Saud University-Computer and Information Sciences* **35**(10), 101810 (2023)

4. Billingsley, G., Dietlmeier, J., Narayanaswamy, V., Spanias, A., O'Connor, N.E.: An L2-normalized spatial attention network for accurate and fast classification of brain tumors in 2D T1-weighted CE-MRI images. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1895–1899 (2023)
5. Bodapati, J.D., Balaaji, B.B.: TumorAwareNet: Deep representation learning with attention based sparse convolutional denoising autoencoder for brain tumor recognition. *Multimedia Tools and Applications* pp. 1–19 (2023)
6. Celik, M., Inik, O.: Development of hybrid models based on deep learning and optimized machine learning algorithms for brain tumor multi-classification. *Expert Syst. Appl.* **238**, 122159 (2024)
7. Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE* **10**, e0140381 (2015)
8. Deepak, S., Ameer, P.: Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **111**, 103345 (2019)
9. Dutta, T.K., Nayak, D.R., Zhang, Y.D.: ARM-Net: Attention-guided residual multiscale cnn for multiclass brain tumor classification using mr images. *Biomed. Signal Process. Control* **87**, 105421 (2024)
10. Hammad, M., ElAffendi, M., Ateya, A.A., Abd El-Latif, A.A.: Efficient brain tumor detection with lightweight end-to-end deep learning model. *Cancers* **15**, 2837 (2023)
11. 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)
12. Jaspin, K., Selvan, S.: Multiclass convolutional neural network based classification for the diagnosis of brain MRI images. *Biomed. Signal Process. Control* **82**, 104542 (2023)
13. Nickparvar, M.: Brain tumor MRI dataset. Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/2645886>. (Accessed on 3rd March) (2021)
14. Öksüz, C., Urhan, O., Güllü, M.K.: An integrated convolutional neural network with attention guidance for improved performance of medical image classification. *Neural Computing and Applications* pp. 1–33 (2023)
15. Ozaltin, O., Coskun, O., Yeniay, O., Subasi, A.: A deep learning approach for detecting stroke from brain CT images using oznet. *Bioengineering* **9**(12), 783 (2022)
16. Shahin, A.I., Aly, S., Aly, W.: A novel multi-class brain tumor classification method based on unsupervised PCANet features. *Neural Comput. Appl.* **35**, 11043–11059 (2023)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
18. UmaMaheswaran, S.K., Ahmad, F., Hegde, R., Alwakeel, A.M., Zahra, S.R.: Enhanced non-contrast computed tomography images for early acute stroke detection using machine learning approach. *Expert Syst. Appl.* **240**, 122559 (2024)
19. 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)
20. Xia, J., Zhou, Y., Tan, L.: DBGGA-Net: Dual branch global-local attention network for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters* (2023)
21. Xiao, Y., Yin, H., Wang, S.H., Zhang, Y.D.: TReC: Transferred ResNet and CBAM for detecting brain diseases. *Front. Neuroinform.* **15**, 781551 (2021)

22. Zhu, H., Wang, W., Ulidowski, I., Zhou, Q., Wang, S., Chen, H., Zhang, Y.: MEEDNets: Medical image classification via ensemble bio-inspired evolutionary DenseNets. *Knowl.-Based Syst.* **280**, 111035 (2023)
23. Zulfiqar, F., Bajwa, U.I., Mehmood, Y.: Multi-class classification of brain tumor types from MR images using efficientnets. *Biomed. Signal Process. Control* **84**, 104777 (2023)



Auxiliary Information Guided Segmentation for the Clinical Target Volume of Cervical Cancer

Shiyun Wang¹  and Yongchao Xu^{1,2}  

¹ School of Computer Science, Wuhan University, 430072 Wuhan, China

{shiyunwang, yongchao.xu}@whu.edu.cn

² Hubei LuoJia Laboratory, 430079 Wuhan, China

Abstract. Segmentation of the clinical target volume (CTV) in cervical cancer is a crucial step for radiotherapy. Existing methods overlook the importance of cancer progression stages and do not consider the spatial relationship between organs at risk (OARs) and the CTV, resulting in suboptimal segmentation outcomes. In this paper, we employ auxiliary information to guide the cervical cancer CTV segmentation. Patient cases are additionally annotated and classified into eight categories based on the cancer progression stages and surgical statuses. These annotations are extended to the size of the inputs and concatenated with them, allowing our network to learn more CTV information from the annotations. We simultaneously train the segmentation of OARs and the CTV, employing a shared encoder and LoRA layers to merge features from OARs and CTV segmentation. By merging the features, the spatial relationship between OARs and the CTV is leveraged. Additionally, we use a Poisson's ratio to model the deformation of OARs under force and implement a data augmentation method by simulating these deformations. Extensive ablation studies and experiments on various baseline networks demonstrate the effectiveness of the proposed method. Our method provides a more generalized and accurate solution for CTV segmentation in cervical cancer.

Keywords: Cervical cancer CTV segmentation · Auxiliary information · Data augmentation

1 Introduction

Cervical cancer ranks as the fourth most frequently diagnosed cancer and is the leading cause of cancer death in many countries [1]. Radiotherapy is currently the main treatment method for cervical cancer [2]. Radiotherapy requires precise delineation of the radiation target area to achieve the best treatment effect without damaging the nearby organs that are at risk for radiation damage.

In a typical radiotherapy setting, doctors use the CTV as the target area for radiation therapy [3]. The CTV contains the gross tumor volume (GTV), which

is the tumor that is visible in images, and a margin for sub-clinical disease spread. Generally, doctors use the GTV as the basis, find a sufficiently large area that covers the maximum invasion range of the cancer without damaging the OARs, and delineate a 3D CTV on a radiotherapy computed tomography (RT-CT) scan. In reality, delineating the CTV for cervical cancer can be particularly challenging, as doctors may need to perform radiotherapy on patients with a resected uterus, or on several areas suspected of cancer cell spread. In these cases, the CTV is no longer just an expansion of the GTV or other visible regions, but a completely virtual volume. The delineation of CTV is highly complex and relies on the experience of doctors [4], making this task a key challenge in radiotherapy planning. This motivates automated approaches to the CTV delineation.

Due to the rapid expansion of computational power and medical data resources, deep learning has been widely applied to various medical image segmentation tasks, in order to facilitate computer-aided diagnosis and intelligent clinical surgery [5]. CTV segmentation based on deep learning is also a widely researched and applied task [6], *e.g.*, prostate cancer CTV segmentation [7,8], esophageal cancer CTV segmentation [9] and rectal cancer CTV segmentation [10]. Unfortunately, due to the significant anatomical differences between cancers in different organs, these approaches often deeply leverage the physiological structural features of GTV and OARs, so their methods cannot be applied to CTV segmentation for different types of cancer. As for the methods for cervical cancer CTV segmentation [11,12], they often delineate a CTV region by simply extending the GTV, which is ineffective when the GTV is not visible due to a resected uterus or when multiple CTVs are present due to the spread of the disease. These previous related works do not consider additional information crucial for CTV segmentation, including the cancer progression stages of patients and the spatial relationship between OARs and CTV.

SAMed [13] (Segment Anything Model for Medical) is built upon the large-scale image segmentation model - Segment Anything Model [14] (SAM), and is a general solution for medical image segmentation that performs remarkably well across various tasks, making it one of the most famous medical image segmentation models. However, directly applying SAMed for the cervical cancer segmentation task is not a promising approach, since this task differs from general medical image segmentation tasks in several ways. First, the cervical cancer CTV does not have any visible boundaries, nor corresponding physiological organs and tissues. Directly performing segmentation on it would result in very low accuracy. Second, even for the dataset collected from the same machine within a similar time period, the shape, extent, and size of CTV can vary greatly among different cases due to different disease stages. Therefore, we need to take these attributes into consideration during the segmentation process. Additionally, the manual slice-by-slice annotations by doctors to obtain the training dataset are time-consuming and labor-intensive. As a result, the training dataset is often insufficient, which means we need to take measures to expand the training data.

Considering these challenges, we propose a novel cervical cancer CTV segmentation method based on the general medical segmentation network SAMed.

The cancer progression stages of the patients affect the spread of cancer cells, thereby influencing the extent of the CTV. In our method, the information about the cancer progression stages is encoded from annotations to generate a map of the same size as the input image, which is then concatenated with the input image to form a new input. Furthermore, the spatial information of OARs effectively guides the network in learning the location of the CTV. We utilize a shared encoder of the SAMed network to obtain features required for both CTV and OARs segmentation. Since the encoder is shared, the features in the LoRA layer undergo fusion and interaction. These features are divided into two parts, used to predict the segmentation of CTV and OARs. We also propose a data augmentation method based on simulating the deformation of OARs. Our method involves modeling the deformation of OARs under external forces using Poisson’s ratio, and simulating various deformations to generate augmented images. Extensive ablation studies and experiments on various baseline networks are performed. Our method achieves 0.8741 *Dice*, 2.4494 *Hd₉₅* and 0.8276 *ASD* based on the SAMed baseline, demonstrating the effectiveness of our proposed method.

We summarize the contributions of this paper as:

- We utilize the relative position relationship between OARs and CTV by fusing features in the LoRA layer.
- We concatenate the surgical statuses and disease stages information with the input to guide the segmentation of CTV.
- We propose a Poisson’s ratio based data augmentation method that simulates deformation of OARs, expanding the training data.
- We evaluate our method on cervical cancer CTV segmentation datasets and achieve better results than the baseline, demonstrating the effectiveness of our method.

2 Related Works

2.1 Medical image segmentation models

Unlike natural image datasets, medical image datasets exhibit relative consistency in attributes such as grayscale, shape, size, and position of foreground instances. Medical image segmentation often leverages the inherent consistency to learn common features. With the rapid advancement in deep learning, the groundbreaking work U-Net [15] emerged. Subsequently, numerous variants of U-Net [16] have been proposed. The mainstream network design strategy is to integrate transformer architectures into the U-Net framework [17,18]. In contrast, SAMed [13] does not require complex network engineering and can perform semantic segmentation on medical images. SAMed applies the low-rank-based [19] (LoRA) fine-tuning strategy to the SAM image encoder, and fine-tunes it together with the mask decoder on labeled medical image segmentation datasets. Not only does SAMed produce excellent segmentation results [20], but its deployment and storage costs are negligible in practical use.

2.2 Clinical target volume segmentation

CTVs lack visible boundaries and do not correspond directly to any specific physiological organs or tissues. Moreover, the shapes, ranges, and sizes of CTVs can vary significantly. Therefore, previous CTV segmentation studies often employ various methods and auxiliary information to improve segmentation performance. Wang et al. [21] use two pipelines to learn the features of OARs masks segmentation and CTV mask segmentation respectively, and design two adapters to adapt the features of OARs pipeline to the CTV segmentation network in both encoder and decoder. Some methods [22, 23] use the GTV, lymph node location and the SDM calculated by OARs masks to guide the segmentation of CTV. Qi et al. [24] use the prior location information of breast cancer to guide the CTV segmentation. Balagopal et al. [25] utilize the style information of physicians to guide the segmentation of post-operative prostate CTV. Based on CT image datasets for cervical cancer patients who received two courses of radiotherapy, Wang et al. [26] use images from both the treatments to align and assist in the segmentation of each other. Some of these methods are tailored to special circumstances. Additionally, some special data are used, *e.g.*, GTV, lymph node and two treatments of the same case. These methods cannot be generalized to CTV segmentation tasks without these relevant data, and their generalization needs to be verified.

3 Method

3.1 Overview

The image of one patient case in the training set is a three-dimensional CT image $\mathbf{x} \in \mathbb{R}^{H \times W \times Z}$ where the resolution of one slice is $H \times W$ and the total number of slices is Z . Our task is to predict its corresponding segmentation map \mathbf{P} with resolution $H \times W \times Z$, where each pixel belongs to an element in a predefined class list $\mathbf{Y} = \{\mathbf{y}_{back}, \mathbf{y}_{ctv}\}$ as close to the ground truth \mathbf{G} as possible. We regard \mathbf{y}_{back} as the background class and \mathbf{y}_{ctv} as the CTV class. We use SAMed as a baseline, freeze all the parameters in the image encoder, and train the trainable bypass in each transformer block.

In this paper, we address challenges in the cervical cancer CTV segmentation task and propose corresponding improvements to SAMed [13]. The pipeline of the proposed method is illustrated in Fig. 1. 1) We utilize the relative positional relationship between OARs and the CTV, addressing the challenge of the absence of visible boundaries in CTV segmentation. 2) To account for differences in disease stages among patient cases, we incorporate past surgeries and disease stage information to guide the network in segmenting CTV masks. The obtained CTV masks are corresponding to the characteristics of the current patient case in terms of the shape, range, and size of ground truth masks. 3) To address the problem of insufficient data, we propose a Poisson’s ratio based data augmentation method based on the simulation of the porous elastic mechanical properties of organs, thereby increasing the amount of training data.

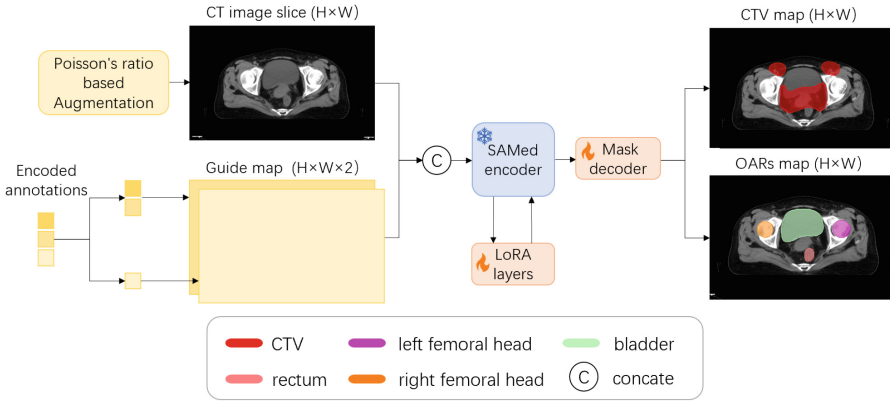


Fig. 1. Pipeline of our proposed method. The guide map is derived from the disease stage information, and it is concatenated with the input CT image to form a new input. Mask decoder predicts multiple segmentation masks, and each mask represents one class in OARs or CTV.

3.2 Utilization of the OARs masks

According to the NCCN (National Comprehensive Cancer Network) guidelines for cervical cancer [27], the positions of the OARs and CTV in cervical cancer radiotherapy planning are closely related. Therefore, an intuitive idea is to use the positional information of the OARs masks to guide the segmentation of the CTV mask. In this paper, we extract the same features from CT images using the encoder of SAMed, then utilize the features to simultaneously segment OARs and CTV masks. OARs masks help the LoRA layer optimize the parameters. The information of the OARs is shared to the CTV segmentation through the LoRA layer.

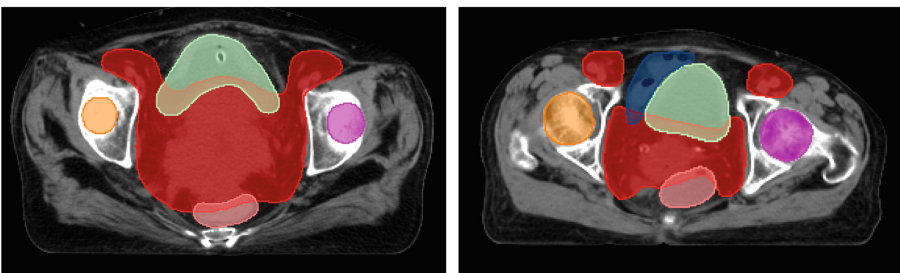


Fig. 2. Visualization of different region masks in two slices of the CT image with overlapping regions between the OARs masks (bladder in green color, rectum in pink color, intestinal tube in blue color), and CTV mask in red color.

As shown in Fig. 2, there are overlapping regions between the masks of OARs and CTV for cervical cancer, which results in CTV not being considered as a new category of the same layer as OARs. The overlapping pixels are both OAR and CTV, leading to annotation conflicts. In the dataset we use, there are C classes of OARs. Consequently, we do not directly consider the CTV as a new category in a $C+1$ classes semantic segmentation task. Instead, the feature maps obtained from the decoder are separated along the channel dimension, resulting in $\mathbf{M}_{oars} \in \mathbb{R}^{C \times H \times W}$ and $\mathbf{M}_{ctv} \in \mathbb{R}^{1 \times H \times W}$ feature maps. The \mathbf{M}_{oars} features are then used to predict the OARs masks \mathbf{P}_{oars} through an argmax operation as:

$$\mathbf{P}_{oars} = \arg \max(\mathbf{M}_{oars}, dim = 0), \quad (1)$$

where $dim = 0$ indicates the argmax operation is performed across the first dimension, with the corresponding loss function defined as:

$$\mathcal{L}_{oar} = (1 - \alpha) \cdot CE(\mathbf{P}_{oars}, \mathbf{G}_{oars}) + \alpha \cdot DICE(\mathbf{P}_{oars}, \mathbf{G}_{oars}), \quad (2)$$

where CE and $DICE$ denote the cross-entropy loss and Dice loss, respectively. The \mathbf{M}_{ctv} features are used to predict the CTV mask by a sigmoid operation as:

$$\mathbf{P}_{ctv} = \begin{cases} 0, & \text{if } Sigmoid(\mathbf{M}_{ctv}) < 0.5, \\ 1, & \text{if } Sigmoid(\mathbf{M}_{ctv}) \geq 0.5, \end{cases} \quad (3)$$

where with the loss function defined as:

$$\mathcal{L}_{ctv} = DICE(\mathbf{P}_{ctv}, \mathbf{G}_{ctv}). \quad (4)$$

With the parameters in our experiments set as $\alpha = 0.8$ and $\beta = 0.7$, the overall loss function can be described as:

$$\mathcal{L} = (1 - \beta) \cdot \mathcal{L}_{oars} + \beta \cdot \mathcal{L}_{ctv}. \quad (5)$$

3.3 Disease Stage Guidance

Under the guidance of multiple specialist oncology clinicians and the NCCN guidelines for cervical cancer [27], a significant factor influencing the CTV mask in cervical cancer is the stage of the cancer disease. For early stage patients, the CTV mask typically includes only the pelvic cavity region with pelvic lymph nodes. In contrast, for patients in the middle or later stages of the disease, where cancer cells have spread, the CTV mask often needs to encompass the inguinal lymph nodes or retroperitoneal lymph nodes. The disease stage cannot be determined solely based on CT images.

In clinical treatment, clinicians often make a comprehensive judgment by integrating other diagnostic information. As a network that assists oncologists, the proposed method uses the disease stage information of patients to guide CTV mask segmentation. We first classify and encode patient cases based on whether the uterus has been resected, whether the cancer has spread to the inguinal lymph nodes, and whether it has spread to the retroperitoneal lymph

nodes. The encoded information is expanded to the same size as the input image $\mathbf{I}_{origin} \in \mathbb{R}^{H \times W}$, resulting in an eight-class map \mathbf{M} . Considering that frozen SAMed encoder can only accept input with three channels, we divide the eight-class map \mathbf{M} into a four-class map \mathbf{M}_{4class} and a two-class map \mathbf{M}_{2class} . The effectiveness of this category separation method is compared in the ablation study. These class maps are then concatenated with the input image to form a new input \mathbf{I}_{new} , as illustrated as:

$$U(m) = \text{Unsqueeze}(m, \text{dim} = 0), \tag{6}$$

$$\mathbf{I}_{new} = \text{Concat}((U(\mathbf{I}_{origin}), U(\mathbf{M}_{4class}), U(\mathbf{M}_{2class})), \text{dim} = 0), \tag{7}$$

where $\text{dim} = 0$ indicates the unsqueeze operation and concatenate operation are performed across the first dimension. The CTV masks range for patient cases with different stages of cancer are shown in Fig. 3. The top and bottom rows of pictures are CT images selected from the dataset. The two rows of pictures are from patient cases with different disease stages, resulting in distinct CTV masks.

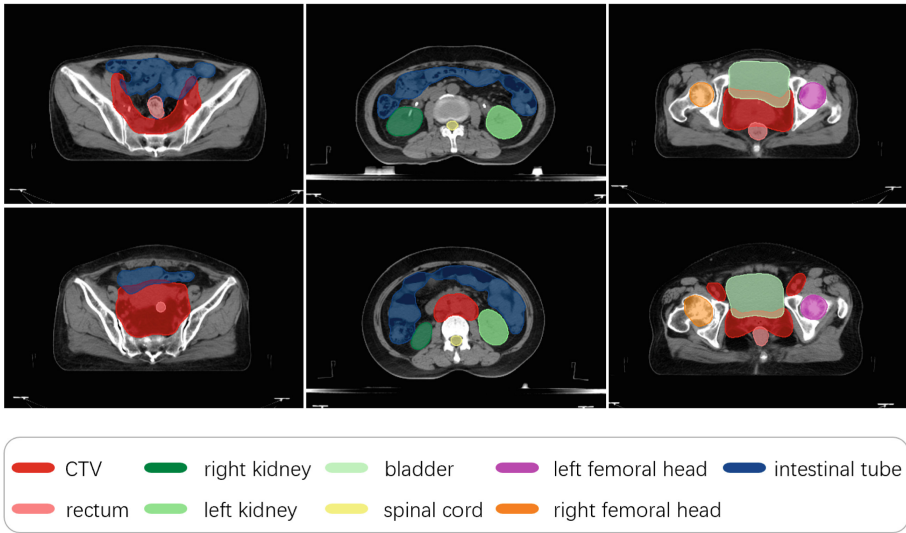


Fig. 3. Visualization of CTVs in cases at different stages of the disease. It can be observed that the CTVs of these different cases vary in terms of number, size, and shape. The top row, from left to right, is from patient cases with a resected uterus, with no spread to retroperitoneal lymph nodes, and with no spread to inguinal lymph nodes. The bottom row is from patient cases with an unresected uterus, spread to retroperitoneal lymph nodes, and spread to inguinal lymph nodes.

3.4 Poisson's Ratio based Data Augmentation

The well-labeled patient cases for cervical cancer CTV segmentation are extremely rare, and the labeling work is cost consuming. Therefore, we achieve data augmentation by simulating the deformation of OARs, thereby expanding the training dataset. The Poisson's ratio [28] is one of the most crucial parameters describing the deformation of materials in three-dimensional space. It represents the ratio of transverse strain to longitudinal strain under compression or uniaxial tension. Using ν to indicate an elastic constant of the material, called the Poisson's ratio. During the elastic deformation stage of the material, the relationship exists between the transverse strain ε_x and the longitudinal strain ε_y is as follow:

$$\varepsilon_x = -\nu\varepsilon_y. \quad (8)$$

Human organs can be considered a porous solid-liquid mixture [29], and the review [30] reported the mechanical properties of whole-body soft tissues, including the Poisson's ratio for human organs in the cervical cancer OARs. We are inspired by anatomy-informed data augmentation [31] calculating a gradient field of OARs masks to represent the deformation direction of OARs. The vector field \mathbf{V} indicates the gradient of the OARs masks \mathbf{M}_{oars} after using Gaussian kernel \mathbf{G}_δ for blurring. Therefore, the vector field \mathbf{V} can be calculated as:

$$\mathbf{V} = \nabla(\mathbf{G}_\delta * \mathbf{M}_{oars}(x, y)). \quad (9)$$

Let v_x and v_y represent the the impact of external forces modeled with Poisson's ratio, \mathbf{I}_{origin} represent the two-dimensional original image slice, and \mathbf{I}_{aug} represent the augmented image slice. The process of data augmentation can be expressed as:

$$\mathbf{I}_{aug}(x, y) = \mathbf{I}_{origin}(x + v_x\mathbf{V}_x(x, y), y + v_y\mathbf{V}_y(x, y)). \quad (10)$$

The values of v_x and v_y are strongly related to the directions of forces. Considering the general situation, the most common force to which human organs are subjected is gravity. And when a patient is at rest, most of the time the direction of gravity is perpendicular or parallel to the direction of \mathbf{V}_x . In order to simplify the problem, we only simulate the external forces perpendicular or parallel to the direction of \mathbf{V}_x . Let \mathbf{I}_{sim} represent the simplified augmented image slice, ν represent the Poisson's ratio, and s represent a deformation scalar based on the forces to control deformation amplitude, the simplified augmentation process is as follow:

$$\mathbf{I}_{sim}(x, y) = \mathbf{I}_{origin}(x + s\mathbf{V}_x(x, y), y - s\nu\mathbf{V}_y(x, y)). \quad (11)$$

The proposed Poisson's ratio based data augmentation simulates the deformation of certain OARs under external forces. This method alters the shape of the real soft tissue around the CTV, increasing the quantity of training cases to increase the generalization ability and the robustness of the network. Its lightweight computing requirements enable easy integration into online training. The visualization of the data augmentation effect is shown in Fig. 4.

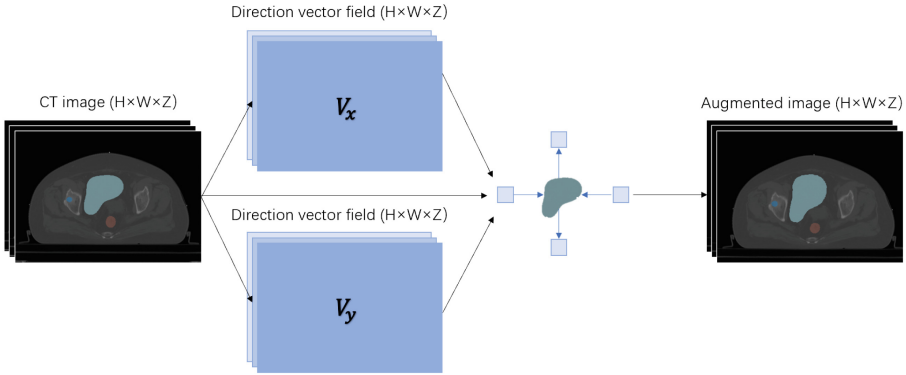


Fig. 4. Visualization results of the Poisson's ratio based data augmentation method. The left images are OARs masks and CT images. We only perform augmentation on OARs. CT images and CTV masks undergo slight change due to the influence of the adjacent OARs.

4 Experiments

4.1 Dataset and preprocessing

The dataset includes CT images and RTstruct files of cervical cancer patient cases obtained from clinical treatments. The collected DICOM format data are converted using the method described in `dcmrtstruct2nii` [33]. The converted image data undergo thorough inspection and verification, with corrections made to address annotation conflicts within the dataset. Poor quality data are filtered out, resulting in a final dataset comprising over 200 cervical cancer cases and approximately 20,000 slices. Each image slice is accompanied by corresponding OARs masks and CTV masks. There are eight organs in OARs masks: bladder, rectum, spinal cord, intestinal tube, left femoral head, right femoral head, left kidney and right kidney. Additionally, data of each case includes extra categorical annotations to represent the impact of different disease stages on the CTV mask range. These annotations primarily cover: whether the uterus was resected, whether the cancer had spread to the inguinal lymph nodes, and whether it had spread to the para-aortic lymph nodes. Specifically, the dataset is ultimately divided randomly into 14,940 two-dimensional slices for the training set and 38 three-dimensional patient cases for the test set.

4.2 Implementation details

In this paper, the model is built in PyTorch and Python3.8 on a server, and an Nvidia RTX 4090 GPU is used for algorithm computations. We adopt the same strategies of data augmentation as SAMed [13]. As for a 512×512 CT image, we input it into SAMed in order to maintain the decent image resolution of the predicted segmentation logits. The output resolution of segmentation logit for

each class is 128×128 , which is smaller than that of UNet-based medical image segmentation models. There are 10 predicted segmentation logits, including one background class, eight OARs classes and one CTV class. As for warmup, we set the initial learning rate I_r to 0.005, the warmup period WP to 250. In the testing stage, input images with sizes of 512×512 are directly fed into the model. For evaluating methods, the dice similarity coefficient (Dice), average symmetric surface distance (ASD) and 95% Hausdorff distance (Hd_{95}) are used to evaluate the performance of the model.

4.3 Main results

Table 1 displays the quantitative evaluation results for the proposed method. These experiments are conducted based on SAMed [13], HiFormer [37], SwinUnet [18], TransUnet [17] and U-Net [15]. We find the proposed methods can benefit these five medical image segmentation models. The comparative experiments are set based on cervical cancer CTV segmentation dataset. And these results prove superiority of the proposed method. Fig.5 shows the qualitative comparisons between our proposed method and SAMed.

Table 1. Main results on the SAMed, HiFormer, SwinUnet, TransUnet and U-Net model on our independent test set. Bl to RK represents the *Dice* of the respective OARs, which include bladder, rectum, spinal cord, intestinal tube, left femoral head, right femoral head, left kidney and right kidney.

Methods	CTV Dice \uparrow	Hd_{95} \downarrow	ASD \downarrow	OARs Avg. Dice \uparrow	Bl \uparrow	Re \uparrow	SC \uparrow	IT \uparrow	LFH \uparrow	RFH \uparrow	LK \uparrow	RK \uparrow
U-Net [15]	0.8076	8.5340	1.8371	0.8283	0.8834	0.8051	0.8179	0.7951	0.7440	0.7311	0.9200	0.9296
U-Net+Ours	0.8361	5.1081	1.4930	0.8354	0.9140	0.8217	0.8772	0.8241	0.6851	0.6692	0.9449	0.9472
TransUnet [17]	0.8103	8.4023	1.7138	0.8305	0.8822	0.7901	0.8172	0.7961	0.7572	0.7525	0.9192	0.9297
TransUnet+Ours	0.8378	4.6014	1.4388	0.8356	0.9023	0.8142	0.8778	0.8139	0.7025	0.6891	0.9416	0.9436
SwinUnet [18]	0.8166	8.0555	1.7235	0.8315	0.9053	0.8137	0.7475	0.8054	0.7662	0.7415	0.9339	0.9386
SwinUnet+Ours	0.8411	4.5538	1.2952	0.8457	0.9212	0.8151	0.8749	0.8307	0.7397	0.6975	0.9391	0.9474
HiFormer [37]	0.8295	5.0354	1.3584	0.8339	0.9089	0.8122	0.7451	0.8129	0.7751	0.7425	0.9347	0.9397
HiFormer+Ours	0.8434	3.5877	1.0523	0.8497	0.9361	0.8256	0.8716	0.8377	0.7161	0.7176	0.94207	0.9506
SAMed [13]	0.8467	3.3166	1.3312	0.8552	0.9360	0.8300	0.8753	0.8407	0.7285	0.7287	0.9422	0.9540
SAMed+Ours	0.8741	2.4494	0.8276	0.8705	0.9417	0.8351	0.7582	0.8233	0.8542	0.8602	0.9436	0.9480

4.4 Ablation study

To validate various design choices of us, we conduct ablation studies on the cervical cancer CTV segmentation tasks, and discuss the details below.

The effect of utilizing the OARs masks. For verifying effectiveness of utilizing the OARs masks, we conducted various experiments. Some experiments use CT images as input for training on SAMed network, only segmenting CTV masks. To leverage the information from OARs masks, the other experiments perform CTV masks segmentation and OARs masks segmentation simultaneously. They split the features conducted from encoder into two feature maps. Each map is used to predict the CTV masks or OARs masks. The information

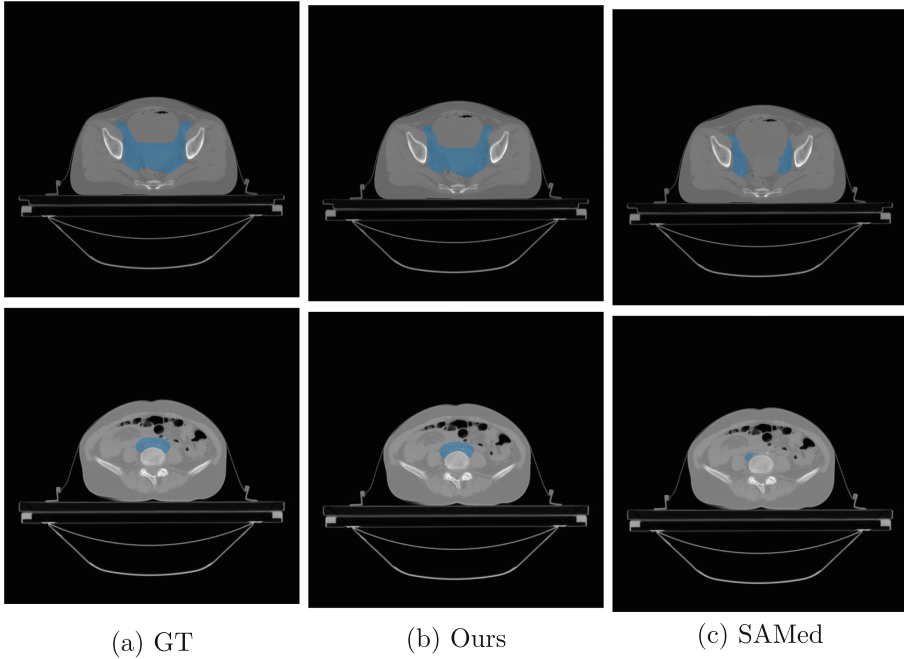


Fig. 5. The comparative analysis of the qualitative aspects between SAMed and our proposed method. Our proposed method integrates the three aforementioned methodologies. The blue regions represent the CTV. As shown in the first row, our method accurately segments the central pelvic area as part of the CTV. The second row demonstrates the precision of our method in segmenting CTV regions affected by cancer cell spread.

of the OARs is shared to the CTV segmentation through the LoRA layer. All of these models are trained with 450 epochs. The experimental results are shown in Table 2.

The Dice scores of training sets with OARs masks in CTV are higher than those of origin training sets, which indicates that using the OARs masks for training is beneficial to improving the segmentation accuracy of cervical cancer CTV mask. In conclusion, utilizing the OARs masks has a promoting effect on the segmentation performance of model.

The effect of disease stages guide. In this part, we primarily evaluate the efficacy of incorporating disease stage information in the customization process of SAMed. As shown in Table 2, the disease stages guide brings significant performance boost for SAMed. We explore the effects of different category combinations on the results. The SAMed encoder is frozen and can only accept input with up to three channels, except for one channel occupied by CT images, additional category information can only use two channels. In the cervical CTV dataset, there are three binary classification categories that can be used. There-

Table 2. Ablation study of our methods in cervical cancer CTV segmentation on SAMed. The first row represents the baseline results. The second to fourth rows display the effectiveness of each individual component. The subsequent rows illustrate the effectiveness of different combinations of components.

OARs	Guidance	Augmentation	Dice \uparrow	Hd ₉₅ \downarrow	ASD \downarrow
			0.8467	3.3166	1.3312
✓			0.8537	3.1690	1.2881
	✓		0.8646	2.6360	1.0271
		✓	0.8597	3.1383	1.1871
✓	✓		0.8655	2.6517	1.0193
✓		✓	0.8610	3.0329	1.1487
	✓	✓	0.8674	2.5043	1.0143
✓	✓	✓	0.8741	2.4494	0.8276

fore, in the experiment, two binary classification maps are combined to form a four-classification map. In Table 3, "Resect & Inguinal" represents the combination of whether the cervical cancer is resected and whether it spreads to inguinal lymph nodes, forming a four-category map. The spread to retroperitoneal lymph nodes is represented as another two-category map. The class-binding rule for the rest of the experiment follows the same naming convention. After evaluating these configurations, we select the best-performing one to use in our method.

Table 3. Ablation study on the combinations of disease stages guidance. "Basic" is pure SAMed without any proposed methods. "Resect & Inguinal" represents the combination of whether cervical cancer resected category and whether the cancer spreads to inguinal lymph nodes category as one channel of SAMed input. Whether the cancer spreads to retroperitoneal lymph nodes is represented as a two-category map and serves as another input channel. The class-binding rule for the rest of the experiment follows the same naming convention.

Network input	Dice \uparrow	Hd ₉₅ \downarrow	ASD \downarrow
Basic	0.8467	3.3166	1.3312
Resect & Inguinal	0.8570	3.0385	1.1653
Resect & Retroperitoneal	0.8579	2.9351	1.1619
Retroperitoneal & Inguinal	0.8646	2.6360	1.0271

The effect of Poisson's ratio based augmentation. The above experiments have proved effectiveness of Poisson's ratio based augmentation. The proposed algorithm is also compared with representative augmentation algorithms based on SAMed on the cervical cancer CTV segmentation task. The data augmentation method based on deformation can be combined with other data augmentation techniques, *e.g.*, translation, rotation, cropping, stitching and color

Table 4. Ablation study on different deformable data augmentation methods. "Basic" is pure SAMed without any proposed methods.

Data augmentation	Dice \uparrow	Hd ₉₅ \downarrow	ASD \downarrow
Basic	0.8467	3.3166	1.3312
Random elastic	0.8390	4.9540	1.3319
Anatomy-informed[31]	0.8528	3.1712	1.2730
Poisson's ratio based	0.8597	3.1383	1.1871

change. Therefore, we only compare results with methods based on deformation. The enhanced results of Poisson's ratio based augmentation and comparison results are shown in Table 4. It shows that the performance of Poisson's ratio based augmentation is better than that of other algorithms with mean Dice scores of 0.8597. The reasons for low accuracies of elastic augmentation and anatomy-informed data augmentation [31] mainly include that the physiological rationality of deformed images is not considered. Thus, the physiological shape and size of CTV in the deformed image are destroyed. Anatomy-informed data augmentation [31] should only be applied to organs such as the rectum or bladder that can undergo significant volume changes over a short period of time. Directly applying it to organs with minimal volume change, may lead to unrealistic expansion and contraction. However, our method simulates the deformation of OARs under external forces, which makes the augmentation more reasonable.

5 Conclusion

In this paper, we address three challenges in the cervical cancer CTV segmentation task and utilize auxiliary information to boost the segmentation performance. We leverage the relative positional relationship between OARs and the CTV to enhance CTV segmentation. Moreover, treatment and disease stage information are incorporated to guide the network in segmenting the cervical cancer CTV. To address the problem of insufficient data, a data augmentation method based on Poisson's ratio is proposed to expand the training data. The effectiveness of our proposed method is validated by experiments conducted on various baseline networks, yielding promising results. Our approach also presents opportunities for further exploration. Currently, our method requires the use of OARs masks. However, in clinical settings where OARs modality data may be unavailable, we can utilize established multi-organ segmentation networks to predict OARs masks. Furthermore, our methodology shows potential for application in the prediction of CTVs for other types of cancer. These possibilities will be explored in future research.

Acknowledgements. This work was supported in part by the National Key Research and Development Program of China (2023YFC2705700), NSFC 6222112, and 62176186, the Innovative Research Group Project of Hubei Province under Grants (2024AFA017).

References

1. 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), 394–424 (2018)
2. Vu, M., Yu, J., Awolude, O.A., Chuang, L.: Cervical cancer worldwide. *Curr. Probl. Cancer* **42**(5), 457–465 (2018)
3. Burnet, N.G., Thomas, S.J., Burton, K.E., Jefferies, S.J.: Defining the tumour and target volumes for radiotherapy. *Cancer Imaging* **4**(2), 153 (2004)
4. Louie, A.V., Rodrigues, G., Olsthoorn, J., Palma, D., Yu, E., Yaremko, B., Ahmad, B., Aivas, I., Gaede, S.: Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. *Radiother. Oncol.* **95**(2), 166–171 (2010)
5. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. *IET Image Proc.* **16**(5), 1243–1267 (2022)
6. Hou, Z., Gao, S., Liu, J., Yin, Y., Zhang, L., Han, Y., Yan, J., Li, S.: Clinical evaluation of deep learning-based automatic clinical target volume segmentation: a single-institution multi-site tumor experience. *Radiol. Med. (Torino)* **128**(10), 1250–1261 (2023)
7. Balagopal, A., Nguyen, D., Morgan, H., Weng, Y., Dohopolski, M., Lin, M.H., Barkousaraie, A.S., Gonzalez, Y., Garant, A., Desai, N., et al.: A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. *Med. Image Anal.* **72**, 102101 (2021)
8. Jin, D., Guo, D., Ho, T.Y., Harrison, A.P., Xiao, J., Tseng, C.K., Lu, L.: Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumors, lymph nodes, and organs at risk. In: *Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention*, pp. 603–612. Springer (2019)
9. Balagopal, A., Morgan, H., Dohopolski, M., Timmerman, R., Shan, J., Heitjan, D.F., Liu, W., Nguyen, D., Hannan, R., Garant, A., et al.: Psa-net: Deep learning-based physician style-aware segmentation network for postoperative prostate cancer clinical target volumes. *Artif. Intell. Med.* **121**, 102195 (2021)
10. Song, Y., Hu, J., Wu, Q., Xu, F., Nie, S., Zhao, Y., Bai, S., Yi, Z.: Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy. *Radiother. Oncol.* **145**, 186–192 (2020)
11. Liu, Z., Liu, X., Guan, H., Zhen, H., Sun, Y., Chen, Q., Chen, Y., Wang, S., Qiu, J.: Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiother. Oncol.* **153**, 172–179 (2020)
12. Zabihollahy, F., Viswanathan, A.N., Schmidt, E.J., Lee, J.: Fully automated segmentation of clinical target volume in cervical cancer from magnetic resonance imaging with convolutional neural network. *J. Appl. Clin. Med. Phys.* **23**(9), e13725 (2022)
13. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. *arXiv preprint [arXiv:2304.13785](https://arxiv.org/abs/2304.13785)* (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: *Proc. of ICCV*. pp. 4015–4026 (2023)

15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention, pp. 234–241. Springer (2015)
16. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. on Medical Imaging* **39**(6), 1856–1867 (2019)
17. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. [arXiv:2102.04306](https://arxiv.org/abs/2102.04306) (2021)
18. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proc. of ECCV, pp. 205–218. Springer (2022)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. [arXiv:2106.09685](https://arxiv.org/abs/2106.09685) (2021)
20. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 108238 (2024)
21. Wang, F., Xu, X., Yang, D., Chen, R.C., Royce, T.J., Wang, A., Lian, J., Lian, C.: Dynamic cross-task representation adaptation for clinical targets co-segmentation in CT image-guided post-prostatectomy radiotherapy. *IEEE Trans. on Medical Imaging* **42**(4), 1046–1055 (2022)
22. Jin, D., Guo, D., Ho, T., Harrison, A.P., Xiao, J., Tseng, C., Lu, L.: Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumors, lymph nodes, and organs at risk. In: Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention, pp. 603–612. Springer (2019)
23. Jin, D., Guo, D., Ho, T., Harrison, A.P., Xiao, J., Tseng, C., Lu, L.: DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. *Med. Image Anal.* **68**, 101909 (2021)
24. Qi, X., Hu, J., Zhang, L., Bai, S., Yi, Z.: Automated segmentation of the clinical target volume in the planning CT for breast cancer using deep neural networks. *IEEE Trans. on Cybernetics* **52**(5), 3446–3456 (2020)
25. Balagopal, A., Morgan, H., Dohopolski, M., Timmerman, R., Shan, J., Heitjan, D.F., Liu, W., Nguyen, D., Hannan, R., Garant, A., et al.: Psa-net: Deep learning-based physician style-aware segmentation network for postoperative prostate cancer clinical target volumes. *Artif. Intell. Med.* **121**, 102195 (2021)
26. Wang, X., Chang, Y., Pei, X., Xu, X.G.: A prior-information-based automatic segmentation method for the clinical target volume in adaptive radiotherapy of cervical cancer. *J. Appl. Clin. Med. Phys.* **25**(5), e14350 (2024)
27. Abu-Rustum, N.R., Yashar, C.M., Bean, S., Bradley, K., Campos, S.M., Chon, H.S., Chu, C., Cohn, D., Crispens, M.A., Damast, S., et al.: NCCN guidelines insights: cervical cancer, version 1.2020: featured updates to the NCCN guidelines. *Journal of the National Comprehensive Cancer Network* **18**(6), 660–666 (2020)
28. Poisson, S.D.: Mémoire sur l'équilibre et le mouvement des corps élastiques. F. Didot (1828)
29. Chaudhuri, O., Cooper-White, J., Janmey, P.A., Mooney, D.J., Shenoy, V.B.: Effects of extracellular matrix viscoelasticity on cellular behaviour. *Nature* **584**(7822), 535–546 (2020)
30. Singh, G., Chanda, A.: Mechanical properties of whole-body soft human tissues: a review. *Biomed. Mater.* **16**(6), 062004 (2021)

31. Kovacs, B., Netzer, N., Baumgartner, M., Eith, C., Bounias, D., Meinzer, C., Jäger, P.F., Zhang, K.S., Floca, R., Schrader, A., et al.: Anatomy-informed data augmentation for enhanced prostate cancer detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 531–540. Springer (2023)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. [arXiv:1711.05101](https://arxiv.org/abs/1711.05101) (2017)
33. Thomas Phil, Thomas Albrecht, Skylar Gay, Mathis Ersted Rasmussen. Sikerdebaard/dcmrtstruct2nii (2023), <https://zenodo.org/records/7705311>
34. Zhou, Z., Qi, L., Shi, Y.: Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In: Proc. of ECCV, pp. 420–436. Springer (2022)
35. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE Trans. on Medical Imaging* **39**(9), 2713–2724 (2020)
36. Liu, Q., Dou, Q., Heng, P.: Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention, pp. 475–485. Springer (2020)
37. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E. K., Cohen-Adad, J., Merhof, D., et al.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proc. of the IEEE/CVF winter conference on applications of computer vision, pp. 6202–6212.(2023)



Synthetic Images with Dense Annotations and Ensemble Learning for DFU Segmentation

Pin Xu¹, Xiongjiang Xiao², Weimin Yuen³, Yanyi Li¹, Kuan Li³,
and Jianping Yin³✉

¹ Dongguan City University, Guangdong, China
pin-xu@outlook.com

² Guangdong University of Science and Technology, Guangdong, China

³ Dongguan University of Technology, Guangdong, China
{yuenwm2022, likuan, jpyin}@dgut.edu.cn

Abstract. Automated image segmentation technology for Diabetic foot ulcer (DFU) management is pivotal in alleviating the healthcare system's workload, considering the severity of DFU as a complication for diabetics. Due to the constraints of annotation costs and privacy, the scale of the publicly available DFU image segmentation datasets is relatively small, which greatly limits the performance improvement of deep learning models. We explore the potential of synthetic image technology in enhancing the performance of DFU image segmentation. We use the FreestyleNet model to generate high-quality synthetic images and employ an error pixel filtering strategy to handle the discrepancies between the synthetic images and masks. To improve the effectiveness and diversity of the synthetic dataset, we specifically designed a mask difficulty calculation method for DFU synthetic images and proposed two innovative resampling strategies based on it. The efficacy of the novel resampling strategies has been demonstrated through comparative experiments conducted against the average sampling method. Furthermore, integrating synthetic image technology with ensemble learning strategies elevates model performance even higher. Our approach achieved a Dice of 73.72% in the Diabetic Foot Ulcer Challenge 2022 on MICCAI 2022, better than the 72.87% Dice that ranked first in the testing phase, ranking second on the Live Leaderboard (as of July 5, 2024). Our code will be released at https://github.com/xupin262/Synthetic_DFU.

Keywords: Image Segmentation · Synthetic Images · Ensemble Learning · DFUC2022

1 Introduction

Diabetic foot ulcer (DFU) is a common and serious complication for diabetes patients, imposing a significant burden on the healthcare system [1, 6, 16]. Accurate automatic segmentation of DFU is crucial for alleviating this burden as it

aids in the development of automated systems for assisted diagnosis and treatment [4, 20].

However, due to limitations such as annotation costs and privacy concerns, the currently available DFU image segmentation datasets are relatively small, significantly restricting the performance improvement of deep learning models. For instance, the DFUC2022 dataset [11] has only 2K labeled images in its training set, and the FUSeg Challenge dataset [29] contains only 1210 labeled images. To address this challenge, we will utilize synthetic image technology to augment the DFU segmentation dataset.

Synthetic Image Technology. Synthetic image technology stems from three primary sources. Firstly, synthetic images from virtual game worlds, commonly used in domain generalization and unsupervised domain adaptation, suffer from significant domain gaps with real images, limiting their effectiveness in fully supervised baselines [21]. Secondly, synthetic images generated through image programs, such as those created through intricate mathematical formulas, have shown promising results when fine-tuned on ImageNet [10]. However, constructing these formulas remains complex and time-consuming, and a performance gap persists between these methods and standard ImageNet pre-trained models in dense prediction tasks. Finally, synthetic images produced by generative models, powerful tools in AI and ML for creating realistic images and videos, occupy a crucial role. This technology is explored in detail below.

Generative models can be categorized into unconditional and conditional types based on their reliance on external information. Unconditional models solely utilize noise as input, whereas conditional models incorporate multimodal data (e.g., labels, text, layout) and necessitate output consistency with the conditional input. Amidst the rapid evolution of generative models, key architectures have emerged as benchmarks, including Generative Adversarial Networks (GANs) [8], Variational Autoencoders (VAE) [13], and Diffusion Models [22, 25]. Furthermore, both Generative Adversarial Networks (GANs) [19, 27] and Diffusion Models [30, 32] can synthesize images based on semantic layouts. This means they create intricate, realistic images from layouts with object position, size, and category. Notably, FreestyleNet [30], proposed by Xue et al., offers high controllability in layout-to-image synthesis. It generates semantic categories beyond training data and allows individual modulation of each category in the text layout, surpassing previous methods and models like Stable Diffusion.

Despite the significant advancements in synthetic image generation technology, its application to specific medical imaging tasks remains fraught with challenges. Brüngel et al. [3] demonstrated this complexity by tripling the DFUC2022 dataset using synthetic images generated through StyleGAN2+ADA and pseudo-labeling via a baseline model ensemble. However, the study highlighted a critical issue: employing a segmentation model, trained on real images, to generate pseudo-masks for synthetic images can introduce annotation errors. As a result, the Dice score paradoxically decreased from 72.11% to 71.69% after the integration of synthetic images. This experimental outcome emphasizes that a mere increase in the quantity of synthetic images does not guarantee enhanced

performance. Quality, not just quantity, is crucial. The following challenges arise in this context: (1) Synthetic images may exhibit artifacts, blurring, or mask mismatches, undermining the efficacy of models trained on such data. Therefore, we introduce an error pixel filtering strategy to ensure that the error pixels in synthetic images are ignored during training. (2) Adding synthetic images to optimized, fully supervised baseline models may have limited marginal performance improvement. To boost performance, we advocate prioritizing 'hard samples' within our methodology. By focusing on resampling these challenging samples, we can unlock significant opportunities for overall performance enhancement.

Ensemble Learning. Ensemble learning is a widely adopted strategy in the field of machine learning, aimed at improving the overall predictive performance by combining the predictions of multiple individual models. The main methods of traditional ensemble learning [18] include Bagging [2], Boosting [7], and Stacking [24].

In modern machine learning, deep learning architectures excel, surpassing traditional shallow models. To harness deep learning's potential and ensemble learning's strengths, ensemble deep learning methods have emerged [17]. These studies integrate ensemble techniques like bagging, boosting, and stacking into various deep learning algorithms[5, 12, 14].

Integrating ensemble learning with deep learning has progressed, but efforts are often task-specific, limiting broader applications [17]. Averaging methods, such as simple and weighted averaging, are commonly used due to their simplicity and broad applicability [9, 17]. Simple averaging averages learner outputs, while weighted averaging assigns weights based on learner performance. When learners perform comparably, simple averaging is effective [23, 26]. However, in diverse learner ensembles, weak or overconfident learners may affect simple averaging's effectiveness [9, 28]. Weighted averaging relies on accurate weight assignment, which can be challenging when learners vary significantly in performance or assessment is difficult. Selecting the right ensemble method is crucial for performance improvement [33].

To tackle the aforementioned challenges, we have meticulously crafted a comprehensive set of methods, each aimed at enhancing the accuracy and robustness of DFU segmentation. The key contributions of this paper are outlined below, illustrating how we have addressed these challenges systematically and innovatively.

- We tailor a set of synthetic dataset processing techniques specifically for the DFU segmentation task, optimizing the quality of the synthetic image dataset and enhancing the generalization ability of the model.
- We design an effective metric to represent the difficulty of DFU masks, and based on this, two innovative DFU synthetic image resampling strategies are proposed.
- We integrate synthetic image technology with ensemble learning strategies and provide a new solution for DFU image processing.

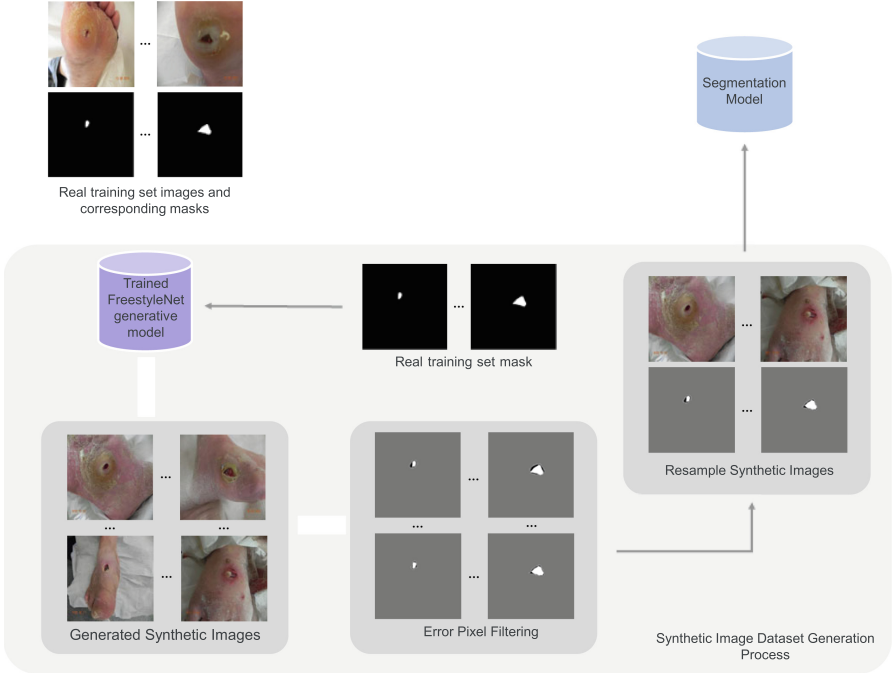


Fig. 1. Illustration of Our Proposed Roadmap

- Our approach achieved a Dice score of 73.72% in the DFUC2022 challenge, better than the 72.87% Dice score achieved by the top-ranked method in the DFUC2022 testing phase. This result highlights the efficacy of our methods.

2 Methods

2.1 Synthesizing Densely Annotated Images

In terms of building and processing synthetic image datasets, this paper has developed a set of specialized processing techniques for the DFU segmentation task. We select the state-of-the-art FreestyleNet[30] model to generate synthetic images and introduce an error pixel filtering strategy aimed at filtering out error pixels that exist between synthetic images and masks. To further enhance the effectiveness and diversity of the synthetic image dataset, this paper specifically designs an effective metric method for DFU synthetic images to quantify the difficulty of the mask. Based on this metric method, two innovative resampling strategies are proposed, and their effectiveness is demonstrated through comparative experiments with the average sampling method. We present an illustration of our pipeline in Figure 1. The following will provide a detailed introduction to the specific methods.

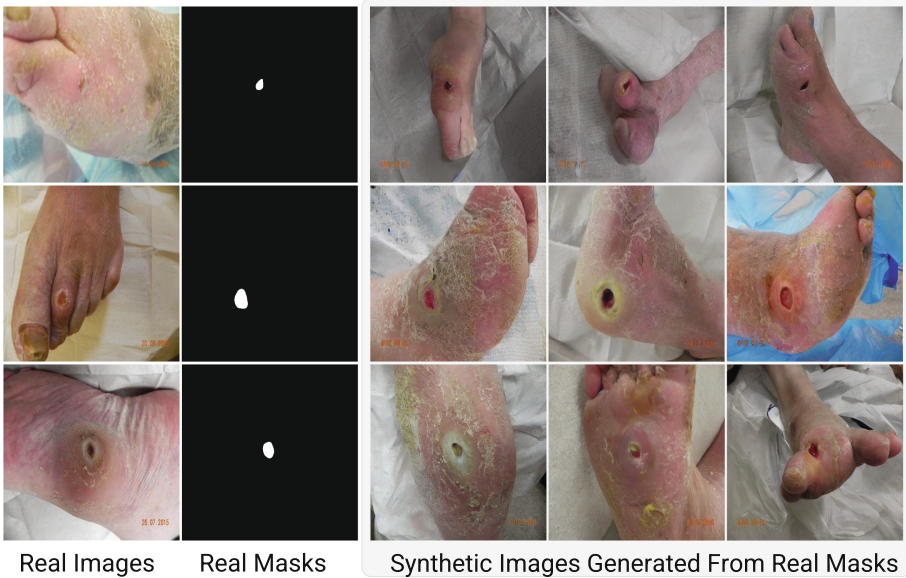


Fig. 2. DFU Synthetic Images Generated by the FreestyleNet Model Based on Real Masks

Generate Synthetic Images We adopt the recent work FreestyleNet to synthesize additional training images based on ground-truth training masks. It is an outstanding generative model that generates realistic images based on semantic layouts. It utilizes real mask labels to synthesize additional training images, shown in Figure 2. If N images are synthesized for each mask, the synthetic data training set will be N times larger than the original set. Each generated image is paired with its corresponding conditional mask during the synthesis process, forming a new training sample with dense annotations. This approach increases the number of training samples and helps reduce annotation costs.

An Effective Difficulty Measurement Method for DFU Masks In images of diabetic foot ulcers, different masks have varying degrees of importance and difficulty. Some masks reflect fewer ulcer areas, with the ulcers being more extensive and having a regular shape. These characteristics enable deep learning models to quickly identify and learn. In sharp contrast, other masks depict a more disordered situation: the edges of the ulcers are complex, and there are frequently many scattered small ulcer areas, which significantly increase the learning difficulty for the model. Figure 3 provides an intuitive display of this. Therefore, how to quantify the difficulty of the mask, and how to accurately sample in the synthetic data training set while considering both the difficulty and diversity of the mask, are the research focuses of this paper. To solve this problem, we propose a novel method to effectively measure the difficulty of DFU masks.

Cross-validation is a commonly used method for model evaluation, and this section innovatively applies it to assess the difficulty of samples in the training set. The five-fold cross-validation method randomly divides the original data into five equally sized subsets. In this process, the model undergoes five rounds of training and validation: each time, one subset is used as the validation set, while the remaining four subsets are merged as the training set. This process is repeated five times to ensure that every sample is validated.

After performing the five-fold cross-validation, the Dice value between the predicted mask and the true mask of DFU can be calculated to obtain the predicted Dice value of each sample in the training set for the segmentation model. Based on these Dice values, it is possible to visually determine which samples have masks that are more difficult to predict correctly (i.e., those with lower Dice values) and which are relatively easier.



Fig. 3. Diversity in DFU Masks

By sorting the Dice values of the training set samples, we can obtain a difficulty ranking for the sample masks, which provides a better understanding of the ease or difficulty of each sample in the segmentation task. Through a thorough analysis of the difficulty distribution of the sample masks, we can gain a deeper understanding of the segmentation difficulty of the samples, providing useful references and guidance for the next step of resampling tasks.

Re-sampling Synthetic Images Based on the Dice difficulty ranking of DFU masks, we propose two resampling strategies for DFU synthetic images: formula-based sampling and limited random sampling. Additionally, we introduce average sampling as a comparative method. We will elaborate on each of these methods.

(1) Formula-based Sampling. Based on the aforementioned DFU mask difficulty measurement method, we have obtained the Dice values for all masks and sorted them in ascending order. In this sorting, masks with lower Dice values indicate greater difficulty, hence we will sample the synthetic images generated by these masks more. The specific number of collections corresponding to each different mask can be calculated using Equation 1.

$$n_d = N_{\max} \cdot \frac{N - d}{N}, \quad (1)$$

where N represents the total number of all real masks. d denotes the Dice ranking position of the real mask (sorted in ascending order), $d \in [0, N - 1]$. N_{\max} is the predefined maximum number of synthetic images for a single real mask. For the mask ranked d -th, n_d represents the final sampling quantity of the synthetic images generated by it, $n_d \in [1, N_{\max}]$.

(2)Limited Random Sampling. The limited random sampling method we proposed is conducted through a series of carefully designed procedures and a random selection process, enabling a rational approach to sampling based on the difficulty level of the masks, thereby improving the model’s generalization ability and performance.

Initially, we randomly eliminate some of the number of synthetic images generated by the mask in proportion to the Dice difficulty ranking of the mask. This step aims to ensure that more challenging samples are gathered during the subsequent sampling process. Specifically, for the masks ranked in the lower 50% by Dice score, which are relatively simple, we retain $0.6N_{\max}$ synthetic images per mask; for the top 50% of the masks, which are considered more difficult, we select the simpler half and retain $0.8N_{\max}$ synthetic images per mask; for the remaining 25% of the most difficult masks, we keep all the generated synthetic images. Through such operations, we ensure a reasonable balance between simple and difficult masks.

Next, we divide the entire range of possible sampling numbers for synthetic images corresponding to a single real mask, $(0, N_{\max}]$, into three equal parts: $(0, N_{\max}/3]$, $(N_{\max}/3, 2N_{\max}/3]$, and $(2N_{\max}/3, N_{\max}]$, and round the boundary values. For all real masks, we randomly select $15\% \pm 2.5\%$ of the total number N , such that the sampling numbers for this portion of the masks fall within the first and last parts $(0, N_{\max}/3]$ and $(2N_{\max}/3, N_{\max}]$. This means that these masks will be chosen with fewer or more synthetic images to ensure comprehensiveness and diversity in the sampling. This step helps to avoid an overemphasis on any particular range of sampling numbers, enabling the model to more comprehensively consider various sampling scenarios during learning, which enhances the model’s robustness and adaptability to changes in sampling numbers.

For the remaining masks (after the above operations, their quantity will be controlled within $70\% \pm 5\%$ of N), we ensure that their sampling numbers fall within the middle part $(N_{\max}/3, 2N_{\max}/3]$. This ensures that the sampling numbers for most masks are neither too many nor too few, maintaining a certain balance. In this sampling method, we need to ensure that the final total number of samples is $N \cdot N_{\max}/2$.

In summary, the Limited Random Sampling Method is an innovative sampling strategy designed for synthetic images. It effectively balances the difficulty level of samples through a carefully crafted sampling mechanism. While maintaining the randomness of the sampling, it ensures that the sample set comprehensively represents the entire data space, which is highly beneficial for enhancing the model’s generalization capabilities.

(3) Average Sampling. To evaluate the impact of sampling strategies on experiments, we established the average sampling as a baseline. This baseline involves collecting a uniform number of synthetic images for each mask.



Fig. 4. Effect diagram of error pixel filtering strategy for DFU synthetic image (in the filtered mask, gray represents the background area, white represents the ulcer area, and black represents the filtered error pixel area)

Filtering Noisy Synthetic Regions Despite the visually realistic nature of synthetic images, they are prone to misleading artifacts. If used indiscriminately for training, these artifacts can significantly disrupt the model, obscuring performance gains, especially with large datasets. To address this, we employed the synthetic image error pixel filtering strategy [31]. This method calculates category-level difficulty scores and filters pixels, effectively eliminating harmful regions that mismatch semantic masks. It aims to adaptively eliminate synthetic areas in the image that do not align well with the corresponding semantic mask. These areas are evaluated using a semantic segmentation model trained on the real dataset, as they often exhibit significant losses. The core idea is to mark synthetic pixels as noise if their loss, predicted by the model for the synthetic image, exceeds the average loss of all pixels of the same category by a certain margin. Noise pixels are then ignored in the loss calculation.

Before implementation, N synthetic images generated by FreestyleNet and a semantic segmentation model trained on the real dataset are required. The strategy involves utilizing the trained segmentation model to compute the dense loss map for the N synthetic images, denoted as $\{\mathbf{L}^i\}_{i=1}^N$ ($\mathbf{L}^i = \{l_k, k \in [1, H \times W]\}$, where l_k represents the pixel-wise loss value, H and W denote the height and width of the map, respectively) for all N synthetic images with their semantic masks $\{\mathbf{M}^i\}_{i=1}^N$. Then, we can calculate the average loss p_j of class j by:

$$p_j = \sum_{i=1}^N \sum_{hw}^{HW} [\mathbb{1}(\mathbf{M}_{hw}^i = j) \times L_{hw}^i] / \sum_{i=1}^N \sum_{hw}^{HW} \mathbb{1}(\mathbf{M}_{hw}^i = j). \quad (2)$$

Where hw indicates the pixel location of (h, w) , $\mathbb{1}(x) = 1$ if x is True, and 0 otherwise.

For a synthetic pixel k with label j , if its loss $l_k > p_j \cdot \alpha$, it is considered potentially noisy. Here, α acts as a tolerance margin. A smaller α results in more synthetic pixels being filtered and excluded from loss computation, enhancing safety. However, an excessively small α may also cause the remaining pixels to lack sufficient information for our model to learn effectively. Fortunately, this filtering strategy exhibits robustness to α , as the performance remains stable across different α values ranging from 1 to 2 [31]. Here α is set to 1.2, and the effect of the error pixel filtering strategy is shown in Figure 2.

Filtering artifacts yields cleaner, more accurate synthetic samples, allowing the model to focus on authentic information. Although simple, this filtering criterion shows great potential. When using only filtered synthetic images for training, the model achieves a Dice score of 69.47% on the DFUC2022 test set, comparable to 72.55% achieved with real images. This highlights its potential for substituting real images, especially in privacy-sensitive medical scenarios.

2.2 Ensemble Learning

We employed ensemble learning methods to improve segmentation performance, including simple and weighted averaging. For the weighted averaging ensemble, we use the following equation to calculate model weights:

$$E(x) = \sum_{i=1}^k w_i M_i, \quad (3)$$

where w_i is the weight for model M_i . Drawing inspiration from Li et al. [15], we calculate the weights using:

$$w_i = \frac{\log\left(\frac{1-d_i}{d_i}\right)}{\sum_{j=1}^k \log\left(\frac{1-d_j}{d_j}\right)}, \quad (4)$$

here, d_i is the Dice score of the i -th model, typically, $d_i \in (0.5, 1)$.

3 Experimental Results and Analysis

Dataset The DFUC2022 dataset [11] is currently the largest publicly available dataset for diabetic foot ulcer segmentation, provided by the organizers of the MICCAI DFUC2022 competition. It features 4,000 high-definition images from the Lancashire Teaching Hospital, with ulcer regions precisely annotated by experts. The dataset encompasses cases of diabetic foot ulcers at various stages of development, with the size of the ulcers ranging from 0.04% to 35.04% of the total image area. These images are divided into a training set and a testing set, each containing 2,000 images. Within the training set, there are 2,304 ulcer cases, over half of which (1,248) have an ulcer area that is less than 1% of the total image area. However, the mask labels of the testing set are not publicly accessible in the DFUC2022 dataset, so participants must submit their results for evaluation and placement in a live leaderboard maintained by the event organizers.

Table 1. Transferability of Synthetic Images to DFUC2022 Real Test Images

Model	Training Data	Dice(%) \uparrow	IoU(%) \uparrow	FNE(%) \downarrow	FPE(%) \downarrow
	Real Synthetic				
Mask2Former (Swin-B)	\checkmark	72.55	62.25	26.36	17.09
	\checkmark	69.47	59.40	20.92	28.86

3.1 Transferability of Synthetic Images to Real Test Images

To comprehensively assess the practical application value of synthetic images, we evaluate the quality of synthetic images through model transfer experiments. Specifically, we test models trained on synthetic images on real images to accurately assess the performance of synthetic images in practical applications.

The filtered and resampled synthetic images are used alone for model training and compared with the corresponding models trained only on real images. As shown in Table 1, in the Mask2Former model with a Swin-B backbone, although the model trained on real data achieved good results (Dice: 72.55%), using only synthetic images for training can also produce a strong model (Dice: 69.47%). The marginal performance gap between the two (Dice: -3.08%) is impressive.

This study validates that synthetic training images, when subjected to filtering and resampling operations, have a significant transfer capability to real test images. This finding suggests they have great potential to replace real training data, especially playing a crucial role in privacy-sensitive medical image segmentation fields.

Table 2. Effectiveness of Filtering and Re-sampling Strategies on DFUC2022

Model	Strategies		Dice(%) \uparrow	IoU(%) \uparrow	FNE(%) \downarrow	FPE(%) \downarrow
	Re-sampling	Filtering				
Mask2Former (Swin-B)	Limited Random		62.42	53.16	27.09	36.48
		\checkmark	69.47	59.40	20.92	28.86
	Formula-based		61.24	51.45	24.58	39.60
		\checkmark	67.33	57.17	19.36	33.85
	Average		48.04	35.97	26.78	55.43
		\checkmark	57.98	49.23	28.17	43.95

3.2 Effectiveness of Filtering and Re-sampling Strategies

This section aims to validate the effectiveness of error pixel filtering methods and resampling strategies. Therefore, a series of ablation experiments were conducted on the DFUC2022 dataset to evaluate these strategies.

Table 2 compares the effects of different filtering and resampling strategies on the model’s transfer performance. Without any special treatment, that is, sampling synthetic images directly with average sampling methods, the Dice score

is only 48.04%. However, after using the error pixel filtering method, the Dice score significantly increased to 57.98%, indicating that the filtering strategy effectively eliminates low-quality pixels and improves the image synthesis quality. In addition, the finite random resampling strategy proposed in this chapter further improved the Dice score to 62.42%, showing effective resampling techniques' important role. When combining the error pixel filtering method comprehensively, the model's transfer performance is maximized, with a Dice score of 69.47%, an improvement of 21.43% over the baseline. The Dice coefficient obtained by formula-based sampling is 67.33%, slightly inferior to finite random sampling methods, but still has significant advantages over average sampling methods.

Experiments show that finite random sampling provides a robust performance improvement, while error pixel filtering methods help in all aspects. In addition, when combining the two, the model shows the best performance in multiple important indicators such as Dice, IoU, and FPE. This result fully demonstrates the synergistic effect of error pixel filtering methods and resampling strategies, and foreshadows their huge potential in improving the quality of synthetic image datasets and their applications in medical image segmentation.

Table 3. Segmentation Performance of Joint Training with Synthetic and Real Images on DFUC2022

Model	Backbone	Training Data		Dice(%) \uparrow	IoU(%) \uparrow	FNE(%) \downarrow	FPE(%) \downarrow	
		Real	Synthetic (Re-sampling)					
			Formula-based					Limited Random
Mask2Former	Swin-T	\checkmark			71.61	61.37	26.87	18.63
		\checkmark	\checkmark		71.95	61.59	25.32	19.54
		\checkmark		\checkmark	72.30	61.98	23.20	21.56
	Swin-S	\checkmark			71.88	61.68	25.03	20.43
		\checkmark	\checkmark		72.07	61.66	24.66	20.32
		\checkmark		\checkmark	72.58	62.37	21.80	22.62
	Swin-B	\checkmark			72.55	62.25	26.36	17.09
		\checkmark	\checkmark		72.89	62.78	21.78	22.05
		\checkmark		\checkmark	72.69	62.54	20.33	23.67
	Swin-L	\checkmark			73.30	63.26	22.52	20.59
		\checkmark	\checkmark		73.58	63.57	21.19	21.25
		\checkmark		\checkmark	73.20	62.94	20.76	22.27
SegFormer	MiT-B2	\checkmark			64.63	54.05	30.43	27.46
		\checkmark	\checkmark		66.43	56.05	26.18	29.21
		\checkmark		\checkmark	67.45	57.17	25.79	27.75
	MiT-B4	\checkmark			65.22	54.45	30.31	26.81
		\checkmark	\checkmark		67.63	57.29	26.18	26.95
		\checkmark		\checkmark	68.39	58.18	25.63	26.17
UPerNet	Swin-L	\checkmark			71.91	61.84	22.32	23.42
		\checkmark	\checkmark		72.28	62.02	23.07	21.85
		\checkmark		\checkmark	71.75	61.74	22.20	24.08
SegNeXt	MSCAN-L	\checkmark			70.96	61.08	23.88	23.65
		\checkmark	\checkmark		71.38	61.33	22.70	23.66
		\checkmark		\checkmark	71.22	61.25	21.97	24.81

3.3 Joint Training of Synthetic and Real Images

The purpose of this section’s experiment is to explore whether combining synthetic images with real images during the training process of deep learning models can improve the performance of the model on the DFUC2022 image segmentation task.

According to Table 3, the joint training of synthetic and real images generally improves the segmentation performance under different model and backbone network combinations. By comparing the results of using only real images for training with those of joint training (synthetic and real images), it can be seen that for the Mask2Former model, regardless of which backbone network is used (Swin-T, Swin-S, Swin-B, Swin-L), the joint training method usually leads to an improvement in Dice score. For example, under the Swin-T backbone, the training Dice score increased from a baseline of 71.61% to 72.30% using the finite random sampling method. This indicates that the introduction of synthetic data has a positive impact on model performance. Further, different synthetic data sampling methods also have varying degrees of impact on performance indicators. For instance, under the Swin-L backbone, using formula-based sampling of synthetic data improved the Dice score from 73.30% with pure real data to 73.58%, while the use of finite random sampling for synthetic data led to a slight decrease to 73.20%. This may mean that different sampling methods have their unique characteristics in aiding model learning. Similar trends were observed in other models such as SegFormer, UPerNet, and SegNeXt, where the joint training of synthetic and real images generally achieved better performance metrics than using real images alone. For example, in the SegFormer model with the MiT-B4 backbone, the Dice score reached 68.39% using the finite random sampling method, significantly outperforming the 65.22% achieved with only real data training.

Overall, the joint training strategy of synthetic and real images has shown a clear advantage in enhancing model segmentation performance. It not only shows improvement in Dice scores but also demonstrates overall performance progress in other metrics such as IoU, FNE, and FPE. By cleverly combining real and synthetic images, the model can learn a richer feature representation, thus possessing better generalization capabilities when dealing with real image data.

Table 4. Segmentation Performance of Ensemble Learning on DFUC2022

Models	Training Data	Ensemble Method	Dice(%) \uparrow	IoU(%) \uparrow	FNE(%) \downarrow	FPE(%) \downarrow
Mask2former (Swin-L)	Real	Simple Averaging	72.50	62.54	21.94	22.74
		Weighted Averaging	72.54	62.58	21.92	22.70
UPerNet (Swin-L)	Real and Synthetic	Simple Averaging	73.66	63.60	20.92	21.57
		Weighted Averaging	73.72	63.64	20.96	21.36

3.4 Ensemble Methods Experiment

To further improve the segmentation capabilities of our models, we investigated the effects of two ensemble techniques—simple averaging and weighted averaging—on the performance when combining two models: Mask2Former (Swin-L) and UPerNet (Swin-L). These two models were trained using real data as well as a combination of real and synthetic data. According to Table 4, when training was conducted with real data only, the weighted averaging method outperformed the simple averaging method across all four metrics: Dice, IoU, FNE, and FPE. Specifically, the Dice coefficient improved to 72.54%, IoU increased to 62.58%, while FNE and FPE decreased to 21.92% and 22.70%, respectively.

When the training data combined both real and synthetic data, the performance of both ensemble methods improved. In this scenario, the weighted averaging method continued to slightly outperform the simple averaging method in terms of Dice and IoU, reaching 73.72% and 63.64%, respectively.

Overall, regardless of whether synthetic data was incorporated into the training or not, the weighted averaging method consistently showed superior performance compared to the simple averaging method. Additionally, training with a combination of real and synthetic data led to better performance across all evaluation metrics. This finding indicates that judiciously selecting ensemble methods and training data is crucial for enhancing model performance.

While model ensemble may consume some Computing resources during the training phase, in practical applications, the simple averaging ensemble and the weighted averaging ensemble methods allow for parallel computation among the models involved in the ensemble. The prediction time only needs to match that of the slowest model. Moreover, the ensemble process involves averaging the model prediction results, which is almost negligible compared to the time required for model prediction. Segmentation accuracy has a crucial impact on diagnostic results for medical images, so further improving segmentation performance through model ensemble has significant value.

4 Conclusions

In this work, we addressed the limited dataset issue in diabetic foot ulcer image segmentation by using synthetic image technology. We tailored a mask difficulty calculation method for DFU synthetic images and designed two resampling methods based on this. Combined with error pixel filtering, we improved the synthetic dataset quality. We further enhanced segmentation performance by integrating synthetic images with ensemble learning. Our methods were validated in the DFUC2022 competition, achieving a Dice score of 73.72%, surpassing the testing leaderboard's top score of 72.87%. This offers valuable insights for other medical image segmentation tasks.

Acknowledgements. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFF0606303, the National Natural Science Foundation of China under Grant 62206054, Research Capacity Enhancement Project of Key Construction Discipline in Guangdong Province under Grant

2022ZDJS028, and the Guangdong Basic and Applied Basic Research Foundation (No.2023B1515120058).

References

1. Armstrong, D.G., Boulton, A.J., Bus, S.A.: Diabetic foot ulcers and their recurrence. *N. Engl. J. Med.* **376**(24), 2367–2375 (2017)
2. Breiman, L.: Bagging predictors. *Machine learning* **24**, 123–140 (1996)
3. Brüngel, R., Koitka, S., Friedrich, C.M.: Unconditionally generated and pseudo-labeled synthetic images for diabetic foot ulcer segmentation dataset extension. In: Yap, M.H., Kendrick, C., Cassidy, B. (eds.) *Diabetic Foot Ulcers Grand Challenge*, vol. 13797, pp. 65–79. Springer International Publishing, Cham (2023)
4. Chae, H.J., Lee, S., Son, H., Han, S., Lim, T.: Generating 3d bio-printable patches using wound segmentation and reconstruction to treat diabetic foot ulcers. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 2539–2549. IEEE (2022)
5. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Real-world image denoising with deep boosting. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(12), 3071–3087 (2019)
6. Edmonds, M., Manu, C., Vas, P.: The current burden of diabetic foot disease. *Clinical Orthopaedics and Trauma* **17**, 88–93 (2021)
7. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: *International Conference on Machine Learning (ICML)*. pp. 148–156. Citeseer, Bari, Italy (1996)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020)
9. Hassib, M., Ali, M., Mohamed, A., Torki, M., Hussein, M.: Diabetic foot ulcer segmentation using convolutional and transformer-based models. In: *Diabetic Foot Ulcers Grand Challenge*, vol. 13797, pp. 83–91. Springer, Cham (2023)
10. Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Zhang, X., Martinez-Noriega, E.J., Inoue, N., Yokota, R.: Replacing labeled real-image datasets with auto-generated contours. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 21232–21241. IEEE, New Orleans, Louisiana, USA (2022)
11. Kendrick, C., Cassidy, B., Pappachan, J.M., O’Shea, C., Fernandez, C.J., Chacko, E., Jacob, K., Reeves, N.D., Yap, M.H.: Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation. [arXiv:2204.11618](https://arxiv.org/abs/2204.11618) (2022)
12. Kim, H.C., Pang, S., Je, H.M., Kim, D., Bang, S.Y.: Support vector machine ensemble with bagging. In: *Pattern Recognition with Support Vector Machines*, vol. 2388, pp. 397–408. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)
13. Kingma, D.P., Welling, M.: Auto-Encoding variational bayes (2022)
14. Li, J., Chang, H., Yang, J.: Sparse deep stacking network for image classification. In: *AAAI Conference on Artificial Intelligence (AAAI)*. vol. 29. AAAI, Austin, Texas USA (2015)
15. Li, K., Yin, J., Lu, Z., Kong, X., Zhang, R., Liu, W.: Multiclass boosting svm using different texture features in hep-2 cell staining pattern classification. In: *International Conference on Pattern Recognition (ICPR)*. pp. 170–173. IEEE (2012)
16. Lo, Z.J., Surendra, N.K., Saxena, A., Car, J.: Clinical and economic burden of diabetic foot ulcers: a 5-year longitudinal multi-ethnic cohort study from the tropics. *Int. Wound J.* **18**(3), 375–386 (2021)

17. Mohammed, A., Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences* **35**(2), 757–774 (2023)
18. Odegua, R.: An empirical study of ensemble techniques (bagging, boosting and stacking). In: *Deep Learning IndabaX*. pp. 1–10. University of Lagos, Abuja, Nigeria (2019)
19. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 2337–2346. IEEE, Long Beach, CA, USA (2019)
20. Ploderer, B., Clark, D., Brown, R., Harman, J., Lazzarini, P.A., Van Netten, J.J.: Self-monitoring diabetes-related foot ulcers with the myfootcare app: A mixed methods study. *Sensors* **23**(5), 2547 (2023)
21. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: *European Conference on Computer Vision (ECCV)*. pp. 102–118. Springer, Amsterdam, Netherlands (2016)
22. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T.: Photorealistic text-to-image diffusion models with deep language understanding. In: *Neural Information Processing Systems (NeurIPS)*. pp. 36479–36494. MIT, New Orleans, USA (2022)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
24. Smyth, P., Wolpert, D.: Stacked density estimation. In: *Neural Information Processing Systems (NeurIPS)*. pp. 668–674. MIT, Denver, USA (1997)
25. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning (ICML)*. pp. 2256–2265. ACM, Lille, France (2015)
26. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 1–9. IEEE, Boston, MA, USA (2015)
27. Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Liu, B., Hua, G., Yu, N.: Diverse semantic image synthesis via probability distribution modeling. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 7962–7971. IEEE, Virtual (2021)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Neural Information Processing Systems (NeurIPS)*. pp. 5998–6008. MIT, Long Beach, USA (2017)
29. Wang, C., Mahbod, A., Ellinger, I., Galdran, A., Gopalakrishnan, S., Niezgoda, J., Yu, Z.: FUSeg: The foot ulcer segmentation challenge (2022)
30. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis (2023)
31. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: FreeMask: Synthetic images with dense annotations make stronger segmentation models. In: *Neural Information Processing Systems (NeurIPS)*. pp. 1–17. MIT (2023)
32. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 3836–3847. IEEE, Paris, France (2023)
33. Zhou, Z.H.: *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, Boca Raton, FL (2012)



SWJEPA: Improving Prostate Cancer Lesion Detection with Shear Wave Elastography and Joint Embedding Predictive Architectures

Markus Bauer¹(✉), Adam Gurwin³, Christoph Augenstein^{1,2}, Bogdan Franczyk², and Bartosz Malkiewicz³

¹ Center for Scalable Data Analytics and Artificial Intelligence, Leipzig, Germany

bauer@wifa.uni-leipzig.de

² Leipzig University, Leipzig, Germany

³ Wroclaw Medical University, Wroclaw, Poland

Abstract. Detecting and localising lesions is a key task in the staging phase of diagnosing and treating prostate cancer (PCa). After a positive digital rectal examination or rise in prostate-specific antigen, pinpointing lesion positions for biopsy using multiparametric magnetic resonance imaging (mpMRI) is crucial. mpMRI and ultrasound (US) imaging already aid in collecting cores for biopsy accurately, a procedure called FBx: mpMRI-targeted US-guided prostate fusion biopsy. Yet, physicians face challenges, e.g., due to limited resolutions of both mpMRI and US. This affects patients' therapy choices, as malignancy assessment accuracy depends on FBx. Recent research aims to improve lesion detection in both mpMRI and US using more objective markers, such as shear wave elastography (SWE). AI can improve FBx using both mpMRI or US data, which has been demonstrated in various studies. However, in the case of mpMRI, labelled lesion examples are still limited, which hinders the performance of state-of-the-art models. Self-supervised learning (SSL) provides a solution by utilising large unannotated databases to create robust feature extractors, enabling the training of case-specific AI models with limited data. Thus, in this paper, we investigate how to improve the models for PCa lesion detection by combining mpMRI and US. We show that recent joint embedding predictive architectures may be a good choice for mpMRI-SSL pretraining. Moreover, we present a false-positive-filtering approach based on real and AI-based SWE, that further improves the mpMRI-model's specificity. Our model achieves state-of-the-art performance of 0.626 average precision in mpMRI-based segmentation and carries the potential to significantly improve lesion detection and localisation accuracy.

Keywords: Artificial Intelligence · Prostate Cancer Detection · Shear Wave Elastography · Magnetic Resonance Imaging

M. Bauer and A. Gurwin—Both authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15327, pp. 359–375, 2025.
https://doi.org/10.1007/978-3-031-78398-2_24

1 Introduction

Prostate cancer (PCa) is the most common solid malignant tumour among men and the second leading cause of cancer-related death worldwide. Fundamental diagnostics include digital rectal examination and prostate specific antigen (PSA) testing, which until recently sufficed for subsequent systematic prostate biopsy (SBx). While SBx is a therapeutic key determinant, reliance on PSA has caused overdiagnosis and unnecessary treatments, with 40-65% surplus biopsies [3] and complications like pain, bleeding, infection, and dysuria [22]. Additionally, SBx misses 30% of clinically significant PCa cases [32].

Recent guidelines recommend multiparametric magnetic resonance imaging (mpMRI) before biopsy [25], shifting towards fusion biopsy (FBx) over SBx. Yet, this diagnostic pathway also has limitations, with higher tumour incidence but lower mortality in biopsied patients [34]. Hence, it's uncertain if FBx reduces overtreatment compared to PSA-driven biopsy. Alternatives like shear wave elastography (SWE) identify lesions by stiffness but need mpMRI to correct false positives from non-cancerous stiff regions, such as calcinosis.

Recent AI advancements enhance mpMRI- and ultrasound (US)-based lesion detection using autoconfiguring segmentation models [17] and are fostered by public datasets [24, 29]. Additionally, Self-supervised learning (SSL) enables training with large, unannotated databases, reducing the need for expert annotation. That way, potent classification and segmentation models can be built without the high burden of initial annotation by experts. This is crucial given the low annotated lesion presence in public mpMRI datasets (e.g., PICA has 15% annotated lesions). SSL therefore carries high potential to improve existing AI approaches.

Thus, in this work, we investigate a joint solution of using both mpMRI and US together with SSL for lesion detection. Our contributions are as follows:

- We compare the joint embedding predictive architecture (JEPA) to other supervised and SSL techniques and show that JEPA is well suited for extracting mpMRI characteristics that enable lesion detection.
- We present first steps towards directly incorporating SWE in AI-lesion-detection.
- We investigate the relation between mpMRI-based AI-lesion-detection and SWE.
- We implement and evaluate an AI-based, generated SWE and prove its predictive value.

2 Related Work

SWE is a vital topic in medical research and is lately backed by guidelines for PCa diagnostics [5]. In the peripheral zone, SWE shows sensitivity and specificity of 96% and 85% respectively [10], and correlates with PCa aggressiveness [19]. Combining SWE with SBx increases biopsy PCa detection rates by 12% [44]. Limited AI research on PCa and SWE exists. One method uses pixel statistics

from SWE regions to train AI models, achieving an area under the receiver operating characteristic (AUROC) of 0.94 for detecting significant PCa [31]. Recent work uses SWE maps with convolutional neural networks, showing comparable AUROC for significant and improved AUROC for insignificant PCa detection [41]. Due to SWE's limited availability, some works demonstrate how to extract it from the US using AI with errors as low as 3.5 kPa [11, 39].

While SWE promises procedural aid in tumour diagnosis, SSL may improve future AI methods for digital cancer staging. Various methods for SSL have been proposed recently that can in particular keep up with the supervised learning pendant [4, 9]. SSL can improve medical AI models, when used as a pre-training technique. Wilson *et al.* [40] achieved accuracies of 69.84 to 81.66% when combining high-frequency ultrasound and histopathology images, whereas an improvement by several percent points over the supervised scenario was observed. Similar results could be achieved for histopathological grading [38], single cell imaging [36], prostate segmentation [13], and brain-tumour anomaly detection techniques [14].

For prostate mpMRI lesion detection, SSL has been proven to be a feasible pretraining strategy [12, 43, 45], with AUROC values of 0.85 ± 0.01 . A limiting factor among all these works, however, appears to be the requirement of a decoder-based architecture, which could potentially limit the learning capabilities. Despite SSL approaches, the supervised training scenario is reported to yield state-of-the-art results. Using a U-Net, dice scores (DSCs) of 0.69 to 0.84 could be achieved [27, 35]. Similarly, nn-UNet [17] achieves a similar DSC of 0.87 [7]. DSCs of 0.79 to 0.84 could furthermore be achieved by other authors, using multiple-level images, attention mechanisms and spatial transforms [21, 33].

3 Methods

3.1 Joint Embedding Predictive Learning

JEPA, a specific SSL branch, is based on the Joint Embedding Architecture (JEA) principle. JEPA employs direct modifications in the latent space, enhancing training speed by eliminating image augmentation, and avoiding decoder-based architectures suspected to limit learning abilities, as highlighted in Assran *et al.*'s IJEPA approach [4]. In JEPA methods, smaller context and larger target patches are selected randomly for each image. The training procedure then shadows context features based on randomly selected positions in the encoder input matrix. The context features are then processed by a learnable attention mask, and forwarded through a predictor network. For the model, the task lies in finding similar latents for shadowed context and original target patches, by interpolating between latents, capturing important data properties. Additionally, a teacher-student configuration with a context and target encoder vision transformer (ViT) [20] is used to avoid mode-collapse.

3.2 Shear Wave Elastography

Utilising acoustic radiation force from multiple focused ultrasound beams, SWE generates shear waves in tissue [2]. These waves' velocity varies with tissue stiffness: higher in stiffer tissues and lower in softer ones. The device generates two shear waves to calculate the Young's modulus (kPa) from their difference, usable for quantitative visualisation.

To collect SWE, avoiding compressing the prostate and rectal wall is crucial, and an elasticity scale of 70-90 kPa is required. Benign prostate tissue in the transitional zone typically measures around 30 kPa to 180 kPa (for metaplasia), while PCa averages 91 kPa. Peripheral and central zones range from 15 to 25 kPa. In the quantitative map, higher values appear in red, lower in blue.

Limitations of SWE include examining large prostates, where gland protrusion can cause compression artefacts, and SWE's 3-4 cm penetration depth may miss anterior lesions [6]. However, most PCa cases are in the accessible peripheral zone [23]. Calcifications can impede performance due to their high stiffness. Furthermore, tissue stiffness varies by imaging plane, with higher values in sagittal compared to axial imaging [28]. Thus, prostate SWE is best performed in the axial plane for representative Young's modulus values.

3.3 Data and Patient Cohort

Internal cohort Internal data were obtained from a single-center, non-randomised prospective study on PCa detection via SWE- and mpMRI-targeted transperineal biopsy at University Hospital in Wroclaw, Poland. Ethical approval was granted with number 129/2023, with written consent from participants. Adult men scheduled for biopsy were eligible if they had prebiopsy 3-T mpMRI with PI-RADS 3+ lesions and PSA determination, and underwent transrectal SWE (Aixplorer® device with the SuperEndocavity SE12-3 probe; tissue stiffness threshold at 90 KPa). The entire prostate was mapped in sagittal sections to identify areas of increased stiffness. Exclusion criteria included prior prostate intervention and clinically spread disease. From June 2023 to February 2024, 133 patients were enrolled; 15 were used for qualitative analysis and testing, and 68 for AI-based SWE training. Each mpMRI included standard T2W, ADC, and DWI. Patients with positive biopsy results were eligible for RARP at University Wroclaw Center of Excellence in Urology. Excised prostates underwent detailed histopathological examination, mapping actual PCa foci onto paper schemes for comparison with mpMRI and SWE images.

The Cancer Imaging Archive For implementation of the lesion-filtering and further validation of the mpMRI-model, we used the MRI-US dataset from the cancer imaging archive (TCIA) [24]. We extracted 277 cases (13 benign; 264 malignant) that contain all of T2W, ADC, DWI and US images as well as masks for lesions and the prostate itself in both mpMRI and US.

PICAI & Prostate158 The PICAI [29] and Prostate158 [1] consist of 1500 (1075 benign; 220 malignant) and 158 (56 benign; 102 malignant) cases of mpMRI with ADC, DWI and T2W. We used them to train and test the mpMRI models. Furthermore, PICAI data was used exclusively for the SSL pretraining.

3.4 Architecture

SWJEPa processes paired cases of mpMRI and corresponding US. For T2W and DWI images, simple scaling from zero to one was performed for each patient. For the ADC images, the mean and standard deviation of each dataset were collected, together with the 99.5 and 0.5% percentiles. Each pixel was then clipped within these percentiles, mean-subtracted and divided by the standard deviation to ensure comparability of different centres' data. Additionally, resampling to a voxel spacing of (0.5, 0.5, 3.0) as suggested by Saha *et al.* [30] was applied.

The SWJEPa pipeline works as follows: First a lesion prediction Z is extracted by a mpMRI-**segmentation model**, that uses an **SSL-pretrained** backbone. Next, an AI-based SWE is created from an **SWE-Model** by feeding a US image that corresponds and is registered to the mpMRI. In the third step, pixels P_z will be sampled from the AI-based SWE, according to the lesion prediction that mainly contain true positives and false positives. In the last step, an SWE-based classifier will **filter false positives lesions** from the predictions Z using the pixels P_z . The whole procedure is depicted in Fig 1 and component details are provided below.

SSL Pretraining For the IJEPa [4] SSL pretraining, we use 775 of the PICAI [29] benign cases. We omitted 300 benign cases and all malignant ones to exclude overfitting effects from the later downstream task in any case. The images were processed as pseudo-RGB (T2W, ADC, DWI) by a ImageNet-pretrained ViT-B16 [20]. We found the default parameters provided by IJEPa [4] sufficient, even though we selected a batch size of 22 (one patient case). Training was done for 300 epochs before passing the weights to the segmentation model, and took about three days on our setup (c.f. Sec. 4).

Segmentation Model After training the IJEPa [4] model for 300 epochs, we froze the weights and extended the model with a U-Net like decoder path. The overall architecture was adopted from Hörst *et al.* [16], whose model was originally designed for cell segmentation and thus is proven to handle small segmentations well. We use a combination of weighted (benign: 0.3; lesion: 1.) cross entropy and dice loss with a learning rate of 0.0001. Training was done according to Sec. 4 and took only a few hours. For the decoder, the first, fourth, seventh, and tenth layer of the ViT were connected via skip connections and a bottleneck size of 256 was used. Furthermore, we use a separate head to predict lesions and the organ itself. We couldn't recognise a difference between training the heads separately and parallel, and thus trained both segmentation outputs in parallel. During inference, we multiply the lesion and prostate predictions, to

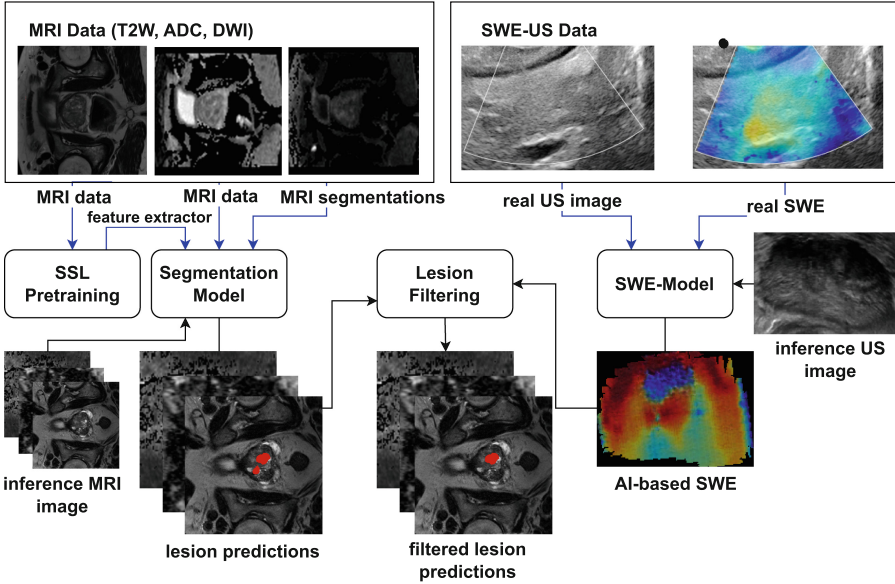


Fig. 1. Overview of the implemented approach. An SSL-pretrained feature extractor is fine-tuned for mpMRI lesion segmentation. Afterwards, corresponding US images are used to create an AI-based estimation of the US-SWE. Using the AI-based SWE, the false positive lesions can be filtered. Blue arrows indicate the training data flow.

mask the actual prostate area. For the segmentation part, however, ADC and DWI channels were set to 0, as they don't show the organ outline necessarily and thus may mislead the model. We also experimented with more recent methods like SwAV [9], but found the performance to be inferior, as using large batches was infeasible due to technical constraints and the size of the dataset. Similarly, using 3D-architectures didn't add value, as for most PICAI solutions [30].

SWE-Model Among the available datasets, only the TCIA [24] offered the required amount of paired cases to train our framework. As the TCIA, however, has no SWE available, we trained a GAN-based pix2pix model [18] with our internal cohort to simulate the SWE in TCIA. We also investigated the performance of the more recent CUT model [26], but found it to create unrealistic predictions, in particular for regions of high Young's modulus. The most recent approach of Dai *et al.* [11] could not immediately be reproduced with our data and thus was postponed to be the subject of future work.

To train the model, first, the SWE regions were extracted, as only a cone inside the whole area had the SWE applied (c.f. Fig. 1 real SWE). Extraction was done by identifying coloured areas based on the image chrominance, cropping the region, and then masking all the non-SWE area to exclude it from training. The model was then trained for a maximum of 300 epochs and using the default

parameters as proposed by Isola *et al.* [18] to replicate the SWE from the raw US. For the lesion filtering experiments, inference was performed using the TCIA US data (c.f. Fig. 5).

Lesion Filtering In the last step of the pipeline, we implemented the SWE-based lesion detection. We used the TCIA [24] dataset for this purpose. As the data contained lesion annotations for the US part, we generated the AI-based SWEs for the dataset and sampled tumour regions from it. Additionally, for each case, we sampled 50 random non-lesion regions of 20×20 pixels. We then proceeded to train a set of machine learning models (c.f. Sec. 4.4 and Tab. 1) on the extracted pixels’ statistics (mean, minimum, maximum) and the labels provided by TCIA [24] (each region was either malignant or benign). Default parameters of Python’s sklearn library were used for training.

This lesion filtering (SWE classifier) was also used in our last experiment, to filter false positive segmentation output of the mpMRI-segmentation model. To pair the AI-based SWE and mpMRI data, the annotated prostate regions were cropped and aligned in both US and mpMRI. Afterwards, the corresponding SWE-pixels of true-positive and false-positive predictions from the mpMRI-classifier were processed by the SWE-classifier.

4 Results

With the components as described in Sec. 3, we’ve built a model that combines mpMRI- and US-based lesion detection. All the models presented below were trained with Python 3.9 and various open-source tools [4, 8, 15, 30], on two CUDA-V100 16 GB GPUs, using early stopping and for at least 100 epochs. Evaluation was done in a five-fold cross validation scenario.

4.1 Qualitative Analysis

Backbone Performance A major part of our study’s intent was to investigate if SSL can be a possible direction to bridge the discrepancy of large data availabilities on the one, but very limited number of annotations on the other hand. The segmentation model uses the IJEPA-pretrained [4] SSL-model as a backbone. To ensure the feature extraction capabilities are not distorted by using the decoder in the actual downstream task (segmentation), first, the raw features were clustered. The result is shown in Fig. 2.

The features extracted by the ViT-based model are high-dimensional as for each case a prediction matrix $M \in \mathbb{R}^{N \times G \times F}$ will be created, whereas N is the number of images per patient case (depth), G is the grid size (16×16 patches) and F is the number of features per patch (768). To properly scatter the features, they first needed to be sampled down further. For that purpose, the mean, standard deviation, and maximum absolute value were calculated along the grid dimension G . Afterwards, the principal component analysis (PCA) was applied to each of the statistics to further downsample the feature size. From the

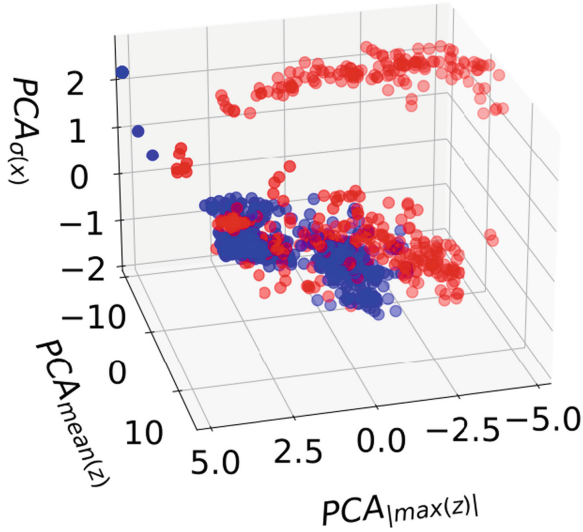


Fig. 2. The IJEPA-based backbone [4] creates features of semantic meaning. When plotting the feature statistics, clearly separable clusters of lesion-infiltrated (red) and lesion-free cases (blue) could be observed.

downsampled features' first PCA main component, a random excerpt was used for plotting. In the resulting scatter plot (c.f. Fig. 2) clearly separable clusters can be found for lesion-infiltrated and lesion-free cases in the upper and lower right half, and in the lower left part respectively. Regarding the statistics, it seems that standard deviation and max absolute value are most predictive. This appears to be plausible, as a lesion-infiltrated case on average will look similar to a lesion-free one due to the lesion only covering a minimal part of the image. The scatter plot therefore suggests that the model captures features of semantic meaning, which is impressive given the fact that only around 17000 images were used for training. Especially among the clusters in the lower half, however, there appears to be a high similarity between images. This, however, is likely of low relevance, as a remaining entanglement is at least expected as a result of the downsampling.

SWE vs. mpMRI In the second part of our qualitative analysis, we investigated the relation between the fine-tuned mpMRI-model's predictions (c.f. Sec. 3.4) and the referring SWE. For this analysis, the 15 cases from our internal cohort were used in a manual that had histopathological confirmation of the lesion positions available as hand-drawing. Generally, the model could be found to handle the cases well but with a lower f_1 score as for the evaluation set taken from the PICAI challenge [30]. While no case was without a predicted lesion, only half of the cases could be found to exclusively match the confirmed positions. For the remaining cases, especially false positive predictions appeared to

be a huge issue, together with some missing sensitivity for small lesions. We therefore investigated the value of the referring SWE data. As the data were recorded in the sagittal plane and smaller misalignments of the prostate could be observed, in a first step we segmented the prostate in the US as described in [37] and manually aligned the images. Afterwards, the axial-converted SWE was plotted together with the mpMRI image and the lesion predictions. In Fig. 3 the predicted lesion lies within a stiff area. This could be advantageous, especially when extracting the biopsies.

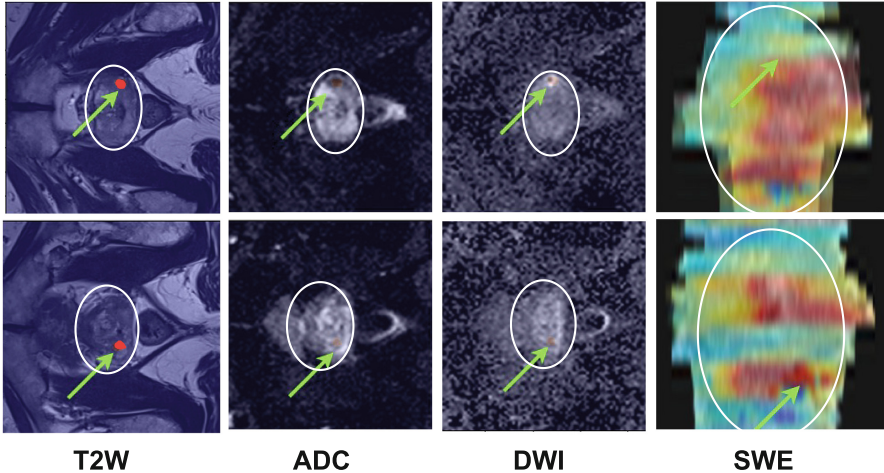


Fig. 3. Sample case of detected MRI-lesion (red) in the prostate (encircled in white) with corresponding SWE. In the SWE image, lower tissue stiffness results in blue and higher stiffness in red colour, whereas the deepest red corresponds to 90 kPa. As indicated by the green arrow, high stiffness is found where AI predicts a lesion.

It can, however, also be observed that stiff regions in SWE exceed the lesion positions. While this could on one hand reveal smaller or less suspicious lesions, SWE is also known to visualise, e.g., calcinosis like a tumour (due to the similar tissue properties) and thus, in terms of usability, relying on both modalities rather than one appears to be advisable from the perspective of the qualitative analysis.

4.2 SSL Downstream Performance

Regarding the mpMRI-model, the qualitative analysis already achieved promising results. To get an in-depth understanding of the model’s performance, we investigated its value as a backbone in a segmentation model. The model performance was then evaluated regarding average precision (AP) and AUROC with the metric as suggested by Saha *et al.* [30] and Bosma *et al.* [8]. We furthermore compared the results against other SSL-based approaches, which are

based on solving the pretext task of restoring subvolumes (3D-patches) of the images and contrastive optimisation. The performance results are depicted in Tab. 1. We resort to comparing only papers using the same metrics as in the PICAI challenge [29] as, e.g., dice score has been proven ineffective in this particular segmentation task (c.f. Yan *et al.* [42]). Similar works that report different metrics can be found in Sec. 3.

Table 1. Performance of the SSL pretrained segmentation-model. The baseline nn-UNet’s lesion level AP could be improved by several percent points, while similar AUROC could be achieved on case level.

Approach	AP	AUROC
nn-UNet [7]	0.593±0.027	0.895±0.008
Subvolume Restoration + Fine-tuning [43]	0.4647±0.0570	0.8624±0.0163
Contrastive SSL + Fine-tuning [45]	0.60±n.A	0.877±n.A.
JEPA + Fine-tuning	0.626±0.068	0.846±0.0476

Our SSL-based model achieves the best AP among the tested approaches. This is important, as AP considers not only if a lesion was detected, but also whether the right position was found. For our biopsy-driven use case, this is a crucial metric, as the case-level AUROC does not reflect if the urologist could successfully sample the tumour. In terms of AUROC, included for completeness, the nn-UNet model reports a higher value. It should, however, be noticed that these results were reported on a private test set rather than the public part of the PICAI data [29]. The internal test sets tended to create more optimistic results [30]. The challenge baseline model, which was also a nn-UNet, achieved inferior results of 0.4556 ± 0.0390 AP and 0.8661 ± 0.0070 AUROC respectively. As the IJEPA [4] model presented here achieved a slightly lower AUROC than the others, we investigated the precision and recall separately. We noticed a significant discrepancy, e.g., for the best-performing model, of 0.860 precision vs. 0.98 recall. This observation supports the findings of our qualitative analysis, as it means the model tends to predict pessimistic (*i.e.*, with a high probability of counting any suspicious pixels as malignant). For our approach that includes a secondary filtering step, we still decided to further build upon the IJEPA [4] model, as an initial pessimistic prediction is a good base for filtering.

In addition to the performance analysis, we also did some further visualisation of successful and failed predictions. They are depicted in Fig. 4.

The visualisation confirmed the findings of both the qualitative and performance analysis once again. While for the example in the upper row both prostate and lesion segmentation are successful (c.f. the corresponding ADC and DWI also), a slight deviation in ADC and DWI, as indicated by the arrow, obviously leads the model to a false-positive prediction.

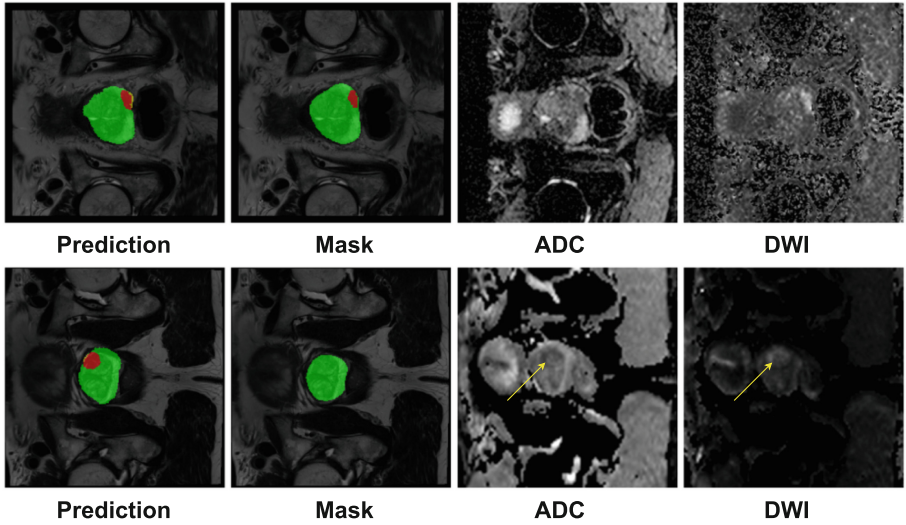


Fig. 4. Visualisation of a successful and failed prediction. While in both rows the prostate segmentation (green) was of high quality (IoU > 0.85) a false positive lesion (red) was detected in the lower row's example. ADC and DWI are non-overlaid to visualise their actual values. The green arrow indicates a clinically insignificant region of high ADC vs. DWI contrast that possibly causes a false positive.

4.3 AI-based SWE

The mpMRI-based model could be found to have a high sensitivity, yet also to predict pessimistic, which can be identified by a remarkably higher amount of false positives compared to false negatives. This could practically foster developments in unnecessary biopsy, which FPx explicitly tries to avoid. As similar effects could be observed for any model tested, we decided to incorporate SWE, as the qualitative analysis and the literature already indicated its value. For the TCIA this required simulation of the SWE, as described in Sec. 3.4. After training on the internal data, our GAN-based model creates AI-based SWEs like in Fig. 5. While inspecting the predictions of 15 cases' images manually, we found most of the regions with high stiffness (c.f. Fig. 5a) to be matching. For images with explicit low stiffness, as in Fig. 5b, false positives could be observed.

As our equipment didn't allow for extracting the raw Young's modulus' data but rather the final quantitative maps, we couldn't directly calculate the real error created by the model. The measured mean squared error in the blue to red chrominance channel of the YCbCr-transformed input and prediction was at $\approx 7.4\%$, which would translate to around 3.5kPa, given the scale of our SWE. This implies that the model's output meets the real SWEs, but is, however, just an estimate. We therefore decided to use the lesion vs. benign prediction performance of the lesion filtering module (c.f. Sec. 3.4 lesion filtering and Tab. 2) as an alternative validation method. It's noteworthy that other models such

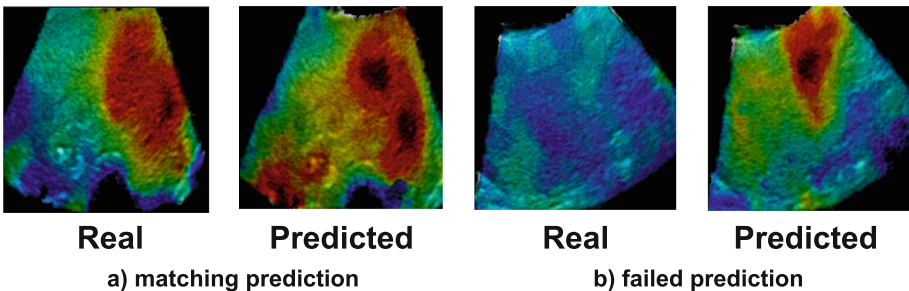
Table 2. Performance of the SWE-based Lesion classifier, for the lesion detection (lesion vs. benign) and false positive filtering (lesion filtering) experiments.

Test Scenario	Model	f_1 score	precision	recall
Lesion vs. Benign	Decision Tree	0.703 ± 0.017	0.712 ± 0.022	0.704 ± 0.018
	Logistic Regression	0.518 ± 0.031	0.735 ± 0.012	0.596 ± 0.019
	SVM	0.791 ± 0.033	0.796 ± 0.036	0.792 ± 0.033
Lesion Filtering	Decision Tree	0.492 ± 0.097	0.584 ± 0.069	0.585 ± 0.081
	Logistic Regression	0.591 ± 0.016	0.594 ± 0.018	0.591 ± 0.012
	SVM	0.578 ± 0.033	0.582 ± 0.036	0.576 ± 0.033

as XGBoost and random forests were also evaluated but not further considered due to their poor tradeoff in performance vs. inference speed. For the lesion vs. benign scenario, the support vector machine (SVM) achieved the best performance regarding f_1 score, precision, and recall. Overall, a high f_1 score of 0.791 ± 0.033 could be achieved. Furthermore, a higher average and max value of the Young’s modulus was found in the lesion region, which meets the expectation.

4.4 Lesion Filtering

In our last experiment, we used the lesion filtering module to filter false positive predictions of the mpMRI-segmentation model. For that purpose, regions where a lesion was detected by the mpMRI-segmentation model were classified by the lesion filtering module and the result was compared to the region’s corresponding label (benign or malignant). We achieved lower f_1 scores than for the initial lesion vs. benign scenario, as depicted in the lesion filtering part of Tab. 2. In contrast to the lesion vs. benign scenario, the logistic regression performed slightly better than the SVM, but both models achieved similar values. We found that in contrast to the mpMRI-model, the US-based one shows less recall but

**Fig. 5.** Examples of the estimated SWE. In figure a) high Young’s modulus values have been correctly captured, while in figure b) stiffness was underestimated.

higher specificity. This is especially the case for detecting true positives, where precision values of 0.644 ± 0.050 were recorded.

Fig. 6 shows three filtered examples. In the top row, the false positive region is below the actual tumour and lies in a low-stiffness area of the AI-based SWE. Hence, it is filtered. For various other cases, we observed that labels don't reach, e.g., deep enough. Therefore, as in the middle row, some lesions may be erroneously marked as false positive and are possibly missed by the lesion filtering model for good reason. We hypothesise that the model performance therefore could even be higher but struggled to determine a matching threshold without unnecessary biasing the model. In the bottom row, the AI prediction covers parts of the tumour, but also larger benign areas, which misleads the lesion filtering.

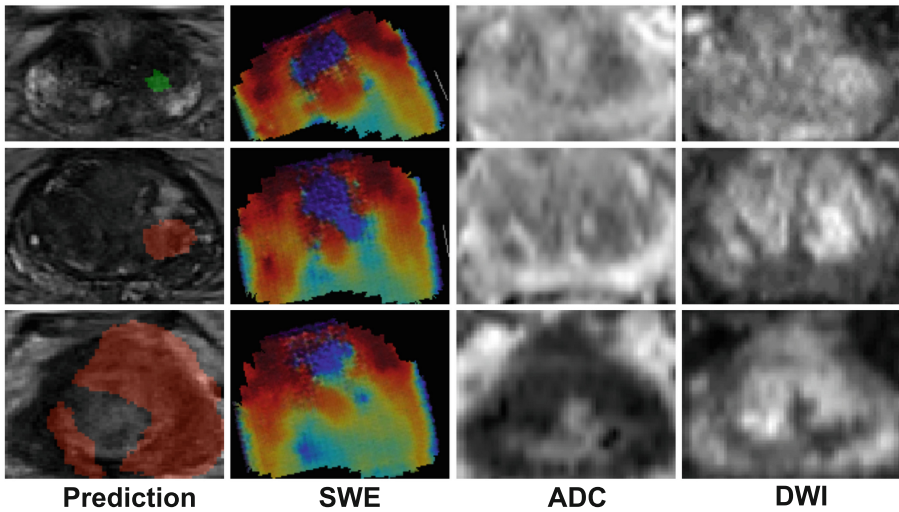


Fig. 6. Examples of the lesion filtering. Top row: correctly identified false positive prediction (green). Middle row: Possibly mislabelled example. Bottom row: Missed false positive prediction.

5 Conclusion

Summary: We developed a method to detect and locate lesions using mpMRI and SWE. Our SWJEPA model achieved higher AP than similar methods and can filter false positives based on SWE. Hence, it has potential for improving FBx and the overall PCa diagnostics pathway, by providing more reliable lesion positions.

The SWJEPA model creates semantically meaningful features for mpMRI images, making it suitable for mpMRI-specific architectures. We anticipate that

pretraining on a larger database, including other tumour types, could yield similar high-quality results as in histopathology by Wang *et al.* [38].

We also implemented and validated an AI-based solution for generating SWE from regular US images, successfully using AI-based SWE maps to identify lesions. This approach could be a viable alternative given the low availability of SWE in hospitals, highlighting SWE's predictive potential.

Limitations: SWJEPA showed a notable decline in performance when tested on data from different hospitals (PICA1 and Prostate158 [1,29]), likely due to differences in DWI/ADC b-values. Additionally, 3D mask sampling and mpMRI image misalignments caused issues, needing better interpolation and registration.

The AI-based SWE showed low deviations from the original, but further investigation is needed with real SWE data, including proper US segmentation for accurate mapping from mpMRI predictions to US.

References

1. Adams, L.C., Makowski, M.R., Engel, G., Rattunde, M., Busch, F., Asbach, P., Niehues, S.M., Vinayahalingam, S., van Ginneken, B., Litjens, G., Bressen, K.K.: Prostate158 - an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. *Comput. Biol. Med.* **148**, 105817 (2022)
2. Ahmad, S., Cao, R., Varghese, T., Bidaut, L., Nabi, G.: Transrectal quantitative shear wave elastography in the detection and characterisation of prostate cancer. *Surg. Endosc.* **27**(9), 3280–3287 (2013)
3. Ahmed, H.U., El-Shater Bosaily, A., Brown, L.C., Gabe, R., Kaplan, R., Parmar, M.K., Collaco-Moraes, Y., Ward, K., Hindley, R.G., Freeman, A., Kirkham, A.P., Oldroyd, R., Parker, C., Emberton, M.: Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389**(10071), 815–822 (2017)
4. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. *CoRR* **abs/2301.08243** (2021)
5. Barr, R.G., Cosgrove, D., Brock, M., Cantisani, V., Correas, J.M., Postema, A.W., Salomon, G., Tsutsumi, M., Xu, H.X., Dietrich, C.F.: Wfumb guidelines and recommendations on the clinical use of ultrasound elastography: Part 5. prostate. *Ultrasound in Medicine & Biology* **43**(1), 27–48 (Jan 2017)
6. Barr, R.G., Memo, R., Schaub, C.R.: Shear wave ultrasound elastography of the prostate: initial results. *Ultrasound Q.* **28**(1), 13–20 (2012)
7. Bosma, J., Peeters, D., Alves, N., Saha, A., Saghir, Z., Jacobs, C., henkjan huisman: Reproducibility of training deep learning models for medical image analysis. In: *Medical Imaging with Deep Learning* (2023), <https://openreview.net/forum?id=MR01DcGST9>
8. Bosma, J.S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., Huisman, H.: Semisupervised learning with report-guided pseudo labels for deep learning-based prostate cancer detection using biparametric mri. *Radiology: Artificial Intelligence* **5**(5) (Sep 2023)
9. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *CoRR* **abs/2006.09882** (2021)





10. Correas, J.M., Tissier, A.M., Khairoune, A., Vassiliu, V., Méjean, A., Hélénon, O., Memo, R., Barr, R.G.: Prostate cancer: Diagnostic performance of real-time shear-wave elastography. *Radiology* **275**(1), 280–289 (2015)
11. Dai, F., Li, Y., Zhu, Y., Li, B., Shi, Q., Chen, Y., Ta, D.: B-mode ultrasound to elastography synthesis using multiscale learning. *Ultrasonics* **138**, 107268 (2024)
12. Fernandez-Quilez, A., Eftestol, T., Kjosavik, S.R., Goodwin, M., Oppedal, K.: Contrasting axial T2W MRI for prostate cancer triage: A self-supervised learning approach. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE (Mar 2022)
13. Fischer, M., Hepp, T., Gatidis, S., Yang, B.: Self-supervised contrastive learning with random walks for medical image segmentation with limited annotations. *Comput. Med. Imaging Graph.* **104**(102174), 102174 (2023)
14. Georgescu, M.I.: Masked Autoencoders for Unsupervised Anomaly Detection in Medical Images. In: Proceedings of KES (2023)
15. Howard, J., Gugger, S.: fastai: A layered api for deep learning. *CoRR abs/2002.04688* (2020)
16. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., Kleesiek, J.: Cellvit: Vision transformers for precise cell segmentation and classification. *CoRR abs/2306.15350* (2023)
17. Isensee, F., Jaeger, P.F., Kohl, S.A.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 (2020)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (2017)
19. Ji, Y., Ruan, L., Ren, W., Dun, G., Liu, J., Zhang, Y., Wan, Q.: Stiffness of prostate gland measured by transrectal real-time shear wave elastography for detection of prostate cancer: a feasibility study. *Br. J. Radiol.* **92**(1097), 20180970 (2019)
20. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. *CoRR abs/2112.13492* (2021)
21. Li, Y., Wu, Y., Huang, M., Zhang, Y., Bai, Z.: Attention-guided multi-scale learning network for automatic prostate and tumor segmentation on MRI. *Comput. Biol. Med.* **165**(107374), 107374 (2023)
22. Loeb, S., Vellekoop, A., Ahmed, H.U., Catto, J., Emberton, M., Nam, R., Rosario, D.J., Scattoni, V., Lotan, Y.: Systematic review of complications of prostate biopsy. *Eur. Urol.* **64**(6), 876–892 (2013)
23. McNeal, J.E.: The zonal anatomy of the prostate. *Prostate* **2**(1), 35–49 (1981)
24. Natarajan, S., Priester, A., Margolis, D., Huang, J., Marks, L.: Prostate MRI and ultrasound with pathology and coordinates of tracked biopsy (prostate-MRI-US-biopsy) (2020)
25. O’Connor, L.P., Lebastchi, A.H., Horuz, R., Rastinehad, A.R., Siddiqui, M.M., Grummet, J., Kastner, C., Ahmed, H.U., Pinto, P.A., Turkbey, B.: Role of multiparametric prostate mri in the management of prostate cancer. *World J. Urol.* **39**(3), 651–659 (2020)
26. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision (2020)
27. Ren, H., Ren, C., Guo, Z., Zhang, G., Luo, X., Ren, Z., Tian, H., Li, W., Yuan, H., Hao, L., Wang, J., Zhang, M.: A novel approach for automatic segmentation of prostate and its lesion regions on magnetic resonance imaging. *Front. Oncol.* **13**, 1095353 (2023)

28. Rouvière, O., Melodelima, C., Hoang Dinh, A., Bratan, F., Pagnoux, G., Sanzalone, T., Crouzet, S., Colombel, M., Mège-Lechevallier, F., Souchon, R.: Stiffness of benign and malignant prostate tissue measured by shear-wave elastography: a preliminary study. *Eur. Radiol.* **27**(5), 1858–1866 (2017)
29. Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H.: The pi-cai challenge: Public training and development dataset (2022)
30. Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H.: Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol) (2022). <https://doi.org/10.5281/zenodo.6667655>
31. Secasan, C.C., Onchis, D., Bardan, R., Cumanas, A., Novacescu, D., Botoca, C., Dema, A., Sporea, I.: Artificial intelligence system for predicting prostate cancer lesions from shear wave elastography measurements. *Curr. Oncol.* **29**(6), 4212–4223 (2022)
32. Serefoglu, E.C., Altinova, S., Ugras, N.S., Akincioglu, E., Asil, E., Balbay, M.D.: How reliable is 12-core prostate biopsy procedure in the detection of prostate cancer? *Can. Urol. Assoc. J.* **7**(5–6), E293–8 (2013)
33. Song, E., Long, J., Ma, G., Liu, H., Hung, C.C., Jin, R., Wang, P., Wang, W.: Prostate lesion segmentation based on a 3D end-to-end convolution neural network with deep multi-scale attention. *Magn. Reson. Imaging* **99**, 98–109 (2023)
34. Stroomberg, H.V., Andersen, M.C., Helgstrand, J.T., Larsen, S.B., Vickers, A.J., Brasso, K., Røder, A.: Standardized prostate cancer incidence and mortality rates following initial nonmalignant biopsy result. *BJU International* **132**(2), 181–187 (Mar 2023)
35. Sun, Z., Wu, P., Cui, Y., Liu, X., Wang, K., Gao, G., Wang, H., Zhang, X., Wang, X.: Deep-learning models for detection and localization of visible clinically significant prostate cancer on multi-parametric MRI. *J. Magn. Reson. Imaging* **58**(4), 1067–1081 (2023)
36. Ternes, L., Dane, M., Gross, S., Labrie, M., Mills, G., Gray, J., Heiser, L., Chang, Y.H.: A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis. *Communications Biology* **5**(1) (2022)
37. Vesal, S., Gayo, I., Bhattacharya, I., Natarajan, S., Marks, L.S., Barratt, D.C., Fan, R.E., Hu, Y., Sonn, G.A., Rusu, M.: Domain generalization for prostate segmentation in transrectal ultrasound images: A multi-center study. *Med. Image Anal.* **82**, 102620 (2022)
38. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* (2022)
39. Wildeboer, R.R., van Sloun, R.J.G., Mannaerts, C.K., Moraes, P.H., Salomon, G., Chammas, M.C., Wijkstra, H., Mischi, M.: Synthetic elastography using b-mode ultrasound through a deep fully convolutional neural network. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **67**(12), 2640–2648 (2020)
40. Wilson, P.F.R., Gilany, M., Jamzad, A., Fooladgar, F., To, M.N.N., Wodlinger, B., Abolmaesumi, P., Mousavi, P.: Self-supervised learning with limited labeled data for prostate cancer detection in high-frequency ultrasound. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **70**(9), 1073–1083 (2023)
41. Wu, H., Fu, J., Ye, H., Zhong, Y., Zhou, X., Zhou, J., Wang, Y.: Multi-modality transrectal ultrasound video classification for identification of clinically significant prostate cancer. *CoRR* **abs/2402.08987** (2024)

42. Yan, W., Yang, Q., Syer, T., Min, Z., Punwani, S., Emberton, M., Barratt, D.C., Chiu, B., Hu, Y.: The impact of using voxel-level segmentation metrics on evaluating multifocal prostate cancer localisation. *CoRR* **abs/2203.16415** (2022)
43. Yuan, Y., Ahn, E., Feng, D., Khadra, M., Kim, J.: Sspt-bpmri: A self-supervised pre-training scheme for improving prostate cancer detection and diagnosis in bi-parametric mri*. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE (Jul 2023)
44. Zhang, M., Wang, P., Yin, B., Fei, X., Xu, X.w., Song, Y.s.: Transrectal shear wave elastography combined with transition zone biopsy for detecting prostate cancer. *Zhonghua nan ke xue = National journal of andrology* **21**(7), 610-614 (July 2015)
45. Zhou, P.: Medical image segmentation with self-supervised learning, https://cs.brown.edu/media/filer_public/35/1c/351ca66e-0628-4d62-b0a1-60cd85897b5c/zhoupeisen.pdf



AdaSVaT: Adaptive Singular Value Thresholding for Adversarial Detection in Fundus Images

Nirmal Joseph¹ , Sudhish N. George¹ , P. M. Ameer¹ ,
and Kiran Raja² 

¹ National Institute of Technology Calicut, Kozhikode, India
{nirmal_p230136ec,sudhish,ameer}@nitc.ac.in

² Norwegian University of Science and Technology Trondheim, Trondheim, Norway
kiran.raja@ntnu.no

Abstract. Existing detection methods for adversarial attacks in medical images mostly rely on prior knowledge about the attacks and the target models. This work introduces a new attack detector termed as AdaSVaT, that is both image and model agnostic, employing specific noise reduction technique through Adaptive Singular Value Thresholding (ASVT). The method exploits the significant impact of adversarial attacks on the lower singular values of an image. The AdaSVaT algorithm adaptively thresholds the singular values of the input image with the help of a linear regressor to generate a low-rank version of the same. Both the original and low-rank versions are experimented with the state-of-the-art classifiers. Adversarial examples are detected by examining the classification inconsistency between the input image and its low-rank version. Additionally, experimental results on three fundus image datasets (Kaggle EyePACS, IDRID and APTOS) prove the negligible loss of information from the images during reconstruction. The proposed method achieves improved adversarial detection accuracy with minimal computational burden while maintaining high structural similarity in both binary and multi-class classification tasks.

Keywords: Adversarial detection · Fundus images · Low rank approximation · Singular value thresholding

1 Introduction

Deep Neural Network (DNN) models are used for medical image classification to reduce the labor-intensive and error-prone nature of the task [1]. Retinal fundus image classification for Diabetic Retinopathy (DR) detection was the first medical image classifier system employed for clinical practice [2]. With the recent advancement in DNN techniques, more models are being incorporated to clinical practices and many of them are outperforming human experts [3]. Despite their efficiency, DNNs in medical image analysis are demonstrably vulnerable to adversarial attacks [4, 5]. These attacks, achieved through imperceptible input

manipulations, lead to misclassifications by DNN models [6]. The ease of crafting such attacks, coupled with the inherent susceptibility of medical images due to their domain-specific features, poses a significant threat [7]. As prior research suggests, adversarial attacks deployed on DNN models can disrupt real-world applications, potentially leading to misdiagnoses and fraudulent claims [8, 9].

Adversarial attacks pose a significant threat to AI-based diabetic retinopathy (DR) detection in clinical settings. By manipulating benign images with imperceptible perturbations, these attacks can induce misclassifications as DR, potentially jeopardizing the patient’s health [22]. Common techniques such as the Fast Gradient Sign Method (FGSM) [26], and Projected Gradient Descent (PGD) [7] exploit the model’s decision boundary to craft adversarial examples that deceive the DNN classifier. These perturbations are visually undetectable by human observers, hindering manual identification. Existing adversarial defense methods, as detailed in Section 1.2, often necessitate significant training on computationally expensive DNN models or operate under restricted white-box assumptions. Consequently, the development of computationally efficient and domain-agnostic detection methods remains an active area of research.

Table 1. Summary of recent adversarial defense methods on fundus images.

Ref.	Defense Type	Task	Attack Modality	Data Modality
[12]	Adversarial Training	Classification	PGD	Fundoscopy
[7]	Adversarial Detection	Classification	FGSM, BIM, PGD, CW	X-ray, Fundoscopy
[13]	Adversarial Training	Classification	PGD, GAP	X-ray, Fundoscopy, Dermoscopy
[14]	Adversarial Training & Feature Enhancement	Classification	FGSM, DeepFool	X-ray, Fundoscopy
[15]	Adversarial Detection, Pre-processing	Classification	FGSM, BIM, PGD, CW, PGD, MI-FGSM	X-ray, Fundoscopy
[16]	Adversarial Detection, Feature Enhancement	Classification	FGSM, PGD, BIM, AutoPGD	X-ray, Fundoscopy
[17]	Adversarial Training, Distillation	Classification	L-BFGS, FGSM	Fundoscopy
[18]	Feature Enhancement	Classification	FGSM	Fundoscopy
[19]	Pre-processing	Segmentation	DAG, I-FGSM	MRI, X-ray, Fundoscopy
[20]	Feature Enhancement	Classification	FGSM, PGD	X-ray, Fundoscopy

This work proposes a new adversarial attack detection method based on Adaptive Singular Value Thresholding (ASVT). The approach leverages low-rank approximations of the input image using the ASVT technique. Classification consistency between the original image and its low-rank counterpart is used to identify potential adversarial examples. The approach is evaluated on three publicly available fundus image datasets.

Rest of the paper is structured as follows: Section 1.1 reviews related work on adversarial image detection. Section 2 details the proposed adversarial detection using adaptive SVT. Section 3 presents experimental evaluations and results, followed by limitations (Section 3.5). The paper concludes in Section 4.

1.1 Related Works

Given the substantial healthcare economy and prevalent medical fraud, extensive research efforts have focused on combating adversarial attacks in medical image analysis. The most common approach is adversarial training, which enhances robustness by using adversarial examples to train the network. [10]. Methods such as feature enhancement and distillation are also in practice to impart adversarial robustness to the DNN model [14, 16, 17]. Adversarial detection is another method in which the adversarial images are identified and avoided from being misclassified [7, 11]. Recent works on adversarial defenses on fundoscopic images are listed in Table 1. The table shows that most adversarial defenses still rely on traditional methods like adversarial training and feature enhancement, with little emphasis on adversarial detection. Traditional methods often suffer from high computational demands and data dependence. Notably, adversarial robustness and adversarial detection serve distinct purposes in machine learning systems. Robust systems prioritize accurate classification of adversarial examples, potentially sacrificing explicit adversarial identification. Conversely, detection methods focus solely on differentiating genuine and adversarial inputs.

Adversarial detection methods are scarce in the medical image domain compared to natural images. Existing methods, while sometimes effective, often struggle with generalizability due to white-box assumptions, limiting their practical application [7, 20, 21]. Even the methods which have reported high accuracy with other datasets (eg: Chest X-rays) were failed to reproduce the same with retinal fundus images [21]. Despite these efforts, developing a generic detection method still remains a challenging task. Hence we have come up with a new and more efficient approach to detect adversarial images in retinal fundus image datasets. Fundoscopic images were chosen as the target modality due to their pioneering role and extensive use in AI-based medical classification models. Furthermore, they present a challenging domain for conventional detection methods, often yielding sub-optimal performance.

1.2 Contributions

Motivated by the gaps identified in the literature, a new approach is introduced for differentiating adversarial images from genuine ones. This work offers significant contributions as listed below.

- A new method for detecting adversarial fundus images using adaptive SVT which works equally good under both binary and multiclass settings. The method exploits the inherent susceptibility of lower singular values to adversarial perturbations, achieving model and image agnostic detection .
- The image entropies are used to calculate the optimum cumulative energy captured by the singular values to perform the thresholding. A low-rank version of the input image is reconstructed by using the thresholded singular values and the adversarial image can be detected if both the predictions on the original and the low-rank versions differ.

- The quality of the reconstructed images is substantiated by comparing their Structural Similarity Index (SSIM) scores with those of the original images. The SSIM scores indicated that the reconstructed images maintained an acceptable level of fidelity to the originals.
- Further, the proposed approach is validated on three publicly available fundus image datasets such as Asia Pacific Tele-Ophthalmology Society (APTOS), Indian Diabetic Retinopathy Image Dataset (IDRID), and Kaggle EyePACS with FGSM and PGD attack methodologies. We validate the proposed approach using the statistical robustness tests.

2 Proposed AdaSVaT Adversarial Detector

The proposed **Adaptive Singular Value Thresholding** technique for adversarial detection in fundus images is termed as AdaSVaT and operates in three stages. Firstly, a pre-trained DNN classifier yields an initial classification for the input image (P1). Secondly, the threshold generation framework determines an adaptive energy threshold for the input image. Subsequently, Singular Value Decomposition (SVD) is performed on the image, and adaptive hard thresholding of the singular values is executed based on the determined energy threshold by considering the cumulative energy captured by the ‘ N ’ most significant singular values. The image’s low-rank version is then reconstructed using the thresholded singular values. Finally, the reconstructed image is classified by the same pre-trained model to get the low-rank prediction (P2). Discrepancies between P1 and P2 indicate an adversarial image. The proposed AdaSVaT adversarial detector is summarized in Fig. 1.

2.1 DNN Model and Attack Settings

ImageNet pre-trained DNN models such as ResNet and INCEPTION are used as the base network for image classification. Transfer learning is employed to achieve high accuracy on the target dataset. FGSM and PGD adversarial attacks are implemented to generate adversarial examples from the original images. These attacks are validated against the DNN model to ensure their effectiveness. The input dataset is created by mixing both attacked and clean images. Images from the input dataset are fed into the classifier to get the original predicted label P1.

2.2 Adaptive SVT and Low-Rank Reconstruction

Prior works in natural image domains treat adversarial perturbations as high-frequency noise and focus on denoising for defense [23–25]. Our approach diverges by exploiting the observation that adversarial attacks predominantly impact the lower singular values compared to the higher ones as shown in Fig. 2. It can be seen from the figure that the adversarial attacks affect the lower singular values to a higher degree as compared to clean images. We base the proposed approach on such an observation to distinguish the singular values of clean versus

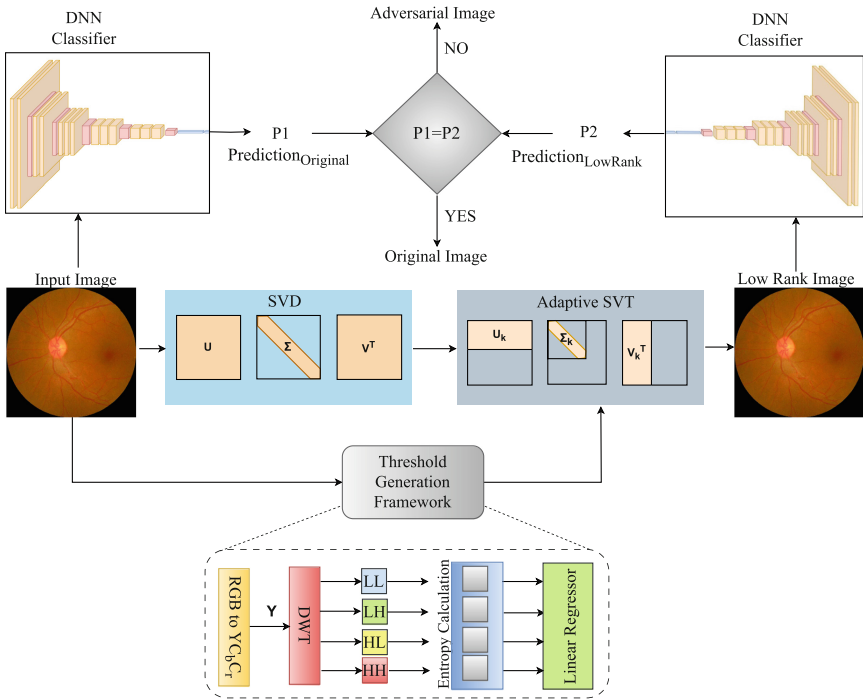


Fig. 1. Overview of the proposed AdaSVaT detector. The detector comprises a DNN classifier and threshold generation framework to perform ASVT. Both the input and its low-rank version after ASVT are fed to the DNN classifier to obtain P1 and P2. Image is detected as adversarial if P1 and P2 differ.

adversarial images. Specifically, the approach is designed by reconstructing the image using only the significant singular values to detect the adversarial images. However, determining a significant percentage of singular values is challenging as it varies from image to image. The proposed approach therefore uses adaptively determined significant singular values (ASVT) instead of conventional SVT.

Threshold Generation Framework: The threshold generation framework calculates the cumulative energy value captured by the significant singular values to perform ASVT. It is achieved by training a linear regressor using entropies derived from the image and corresponding energy value labels as detailed below.

Entropy Calculation: A key challenge lies in the selection of optimal threshold in SVT operation. To maintain the balance between noise reduction and feature preservation, entropy-based adaptive thresholding is employed. Image entropy reflects the degree of adversarial perturbation, with lower values indicating stronger attacks [24]. Consequently, a higher threshold is applied to low-entropy images (significant perturbation) to suppress noise, while high-entropy images (minimal/no perturbation) utilize a lower threshold for optimal fea-

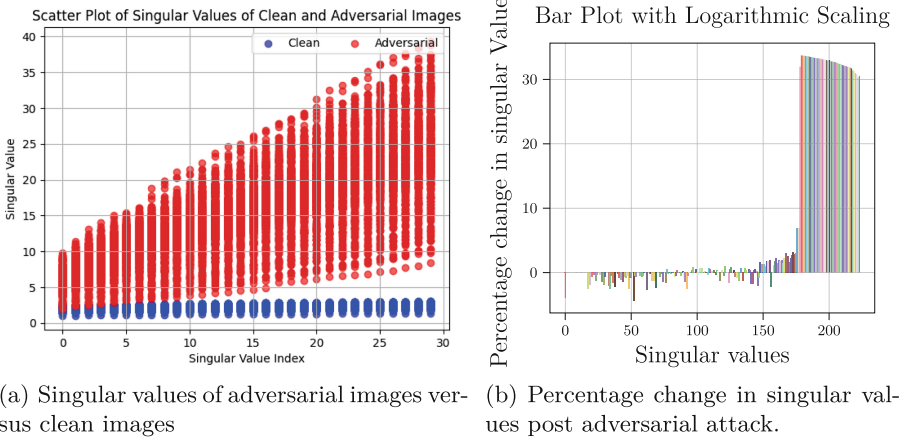


Fig. 2. Plots depicting the impact of adversarial attacks on singular values. (a) The scatter plot presents the lowest 30 singular values before and after the attack for 4000 images (2000 clean and 2000 adversarial). A clear distinction is evident in the lower singular values, which diminishes progressively towards the higher values. (b) Shows the percentage change in singular values post-attack (X-axis corresponds to the indices of these singular values). Notably, lower singular values demonstrate considerable alteration, while the impact on higher singular values is comparatively negligible.

ture retention [24]. Converting an image from \mathcal{RGB} to $\mathcal{YC}_b\mathcal{C}_r$ before performing thresholding enhances precision and accuracy. This approach leverages the advantages of $\mathcal{YC}_b\mathcal{C}_r$ color space (noise or artifacts often appear more distinctly in the luminance component than in chrominance) and wavelet transform to ensure the robustness of the thresholding process. By converting the image to $\mathcal{YC}_b\mathcal{C}_r$, we separate the luminance (\mathcal{Y}) component (represents the brightness) from the chrominance (\mathcal{C}_b and \mathcal{C}_r) components (encode the color information). This separation helps to consolidate the effects of adversarial attacks, as any perturbations introduced across the \mathcal{RGB} channels are integrated into the luminance component. For an 8-bit \mathcal{RGB} input image \mathcal{D} , it can be calculated as

$$\mathcal{D}_{(RGB)} \rightarrow \mathcal{D}_{(\mathcal{YC}_b\mathcal{C}_r)} \quad (1)$$

Subsequently, applying wavelet transform to the luminance component ($\mathcal{Y}(\mathcal{D})$) of the image enables decomposition into four frequency bands: \mathcal{LL} , \mathcal{LH} , \mathcal{HL} , and \mathcal{HH} . Computing entropies for each of these bands provides distinct and finite values, which offer a more reliable basis for thresholding. The Haar wavelet transform (WT) of $\mathcal{Y}(\mathcal{D})$ can be represented as,

$$WT(j, k) = (1/\text{sqrt}(2))^j * \sum_{n=0}^{2^j-1} h_n^{j-1} * \mathcal{Y}(\mathcal{D}_{k,n}) \quad (2)$$

where ‘ j ’ represents the level of decomposition. For $j = 2$, the components are,

$$\mathcal{W}T_{j=2}(\mathcal{Y}(\mathcal{D})) \rightarrow \mathcal{Y}_{LL}, \mathcal{Y}_{LH}, \mathcal{Y}_{HL}, \mathcal{Y}_{HH}$$

Now the entropies can be calculated as,

$$\mathcal{H}(\mathcal{Y}) = - \sum_{i=0}^{L-1} \mathcal{P}(i) \log_2 \mathcal{P}(i) \tag{3}$$

$\mathcal{P}(i)$ represents the probability mass function of pixel intensity values and $L = 8$ for an 8 bit image. Performing the same on all the four bands will yield

$$\mathcal{H}(\mathcal{Y}_{LL}), \mathcal{H}(\mathcal{Y}_{LH}), \mathcal{H}(\mathcal{Y}_{HL}), \text{ and } \mathcal{H}(\mathcal{Y}_{HH})$$

Linear Regressor to Predict Cumulative Energy Value: Image entropy guides the selection of the optimal singular value threshold for low-rank reconstruction. This exploits the inherent property that most image information resides in the higher singular values as shown in Fig. 3. To address entropy-based threshold variations across datasets, a linear regressor, trained on diverse data, predicts the cumulative energy threshold based on the calculated entropy. The singular values of the input image \mathcal{D} of size $m \times n$ can be calculated as,

$$\mathcal{D} = \mathcal{U}\Sigma\mathcal{V}^T \tag{4}$$

where:

- \mathcal{U} is an $m \times m$ orthogonal matrix containing the left singular vectors.
- Σ is an $m \times n$ diagonal matrix with singular values σ_i as the diagonal elements.
- \mathcal{V}^T is an $n \times n$ orthogonal matrix containing the right singular vectors.

Now the normalized singular values ($\hat{\sigma}_i$) are to be added for calculating the cumulative sum (\mathcal{C}_k).

$$\hat{\sigma}_i = \frac{\sigma_i}{\sum_{j=1}^{\min(m,n)} \sigma_j} \tag{5}$$

$$\mathcal{C}_k = \sum_{i=1}^k \hat{\sigma}_i \tag{6}$$

The exact fraction of \mathcal{C}_k necessary for optimal reconstruction, or the cumulative energy threshold ($\hat{\mathcal{C}}_k$) is determined experimentally across various datasets.

Singular Value Thresholding and Low-Rank Reconstruction: Following the generation of the cumulative energy threshold ($\hat{\mathcal{C}}_k$), hard thresholding is applied to the singular values. Only the singular values contributing to the desired cumulative energy are retained for low-rank image reconstruction. The low-rank approximation of the original image \mathcal{D} is obtained by retaining only the k most significant singular values in accordance with their corresponding singular vectors:

$$\mathcal{D}_k = \mathcal{U}_k \Sigma_k \mathcal{V}_k^T \tag{7}$$

where:

- U_k , Σ_k , and V_k^T are the matrices containing the k most significant singular vectors and singular values.
- k is the number of significant singular values, obtained from the number of singular values required to maintain the energy threshold (\hat{C}_k).

After obtaining the reconstructed image, it is classified by using the same classifier. Let the obtained predicted label be P2.

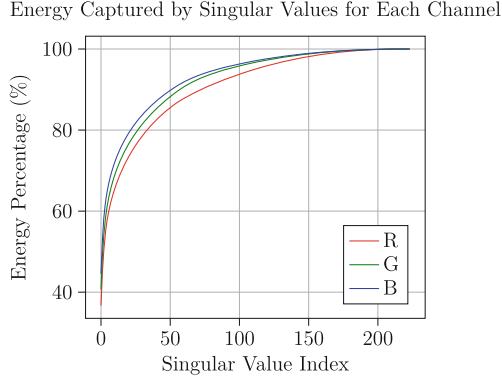


Fig. 3. Cumulative energy captured by the singular values across each channel of a 224×224 image from EyePACS dataset. Most of the energy is concentrated on the higher singular values, indicating their dominant contribution to the overall image representation.

2.3 Attack Detection

The input image is detected as original or adversarial based on the two predicted labels P1 and P2. Since the low-rank version has already removed most of the adversarial noise, if the input image is perturbed, P2 will be different from P1. Conversely, if the input image is not perturbed, P2 will be same as P1, as the reconstruction preserves most of the image-specific features. An algorithmic description of the entire procedure is included in Algorithm 1.

3 Results and Discussions

In this section, we provide a detailed description of the experimental setup, the datasets used and the quantitative and qualitative analysis of the proposed strategy.

3.1 Dataset

Three primary fundoscopic image datasets, Kaggle EyePACS¹, IDRID² and APTOS³, are utilized in this work. These datasets categorize images into five

¹ <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>.

² <http://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset>.

³ <https://www.kaggle.com/c/aptos2019-blindness-detection/data>.

distinct classes: Normal, Mild, Moderate, Severe, and Proliferate. Our study addresses both binary and multi-class classification tasks. For the binary classification, we divided the original dataset into two classes: Normal and Diabetic Retinopathy (DR) by merging all the four abnormal classes into DR. The experiments utilized the entire IDRID and APTOS training set (413 and 2048 images respectively) alongside a random subset of 2000 images drawn from the EyePACS dataset.

Algorithm 1: The AdaSVaT Algorithm

Initialization: pretrained classification model \mathbf{M} and pretrained linear regressor \mathbf{R} .

Input : Images

```

1 for each input image  $\mathcal{D}$  do
2    $P_{\text{Original}} \leftarrow \mathbf{M}(\mathcal{D})$  ; // Classify the input image
3    $\mathcal{D}_{(RGB)} \rightarrow \mathcal{D}_{(\mathcal{Y}\mathcal{C}_b\mathcal{C}_r)}$  ; // Convert from RGB to  $\mathcal{Y}\mathcal{C}_b\mathcal{C}_r$ 
   /* Wavelet transform of the Y component of  $\mathcal{D}_{(\mathcal{Y}\mathcal{C}_b\mathcal{C}_r)}$ . */
4    $WT_{j=2}(\mathcal{Y}(\mathcal{D})) \rightarrow \mathcal{Y}_{LL}, \mathcal{Y}_{LH}, \mathcal{Y}_{HL}, \mathcal{Y}_{HH}$ 
   /* Compute the entropies. */
5    $\mathcal{H}(WT(\mathcal{Y}(\mathcal{D}))) \rightarrow \mathcal{H}(\mathcal{Y}_{LL}), \mathcal{H}(\mathcal{Y}_{LH}), \mathcal{H}(\mathcal{Y}_{HL}), \mathcal{H}(\mathcal{Y}_{HH})$ 
   /* Predict the cumulative energy threshold  $\hat{\mathcal{C}}_k$ . */
6    $\hat{\mathcal{C}}_k \leftarrow \mathbf{R}(\mathcal{H}(\mathcal{Y}_{LL}), \mathcal{H}(\mathcal{Y}_{LH}), \mathcal{H}(\mathcal{Y}_{HL}), \mathcal{H}(\mathcal{Y}_{HH}))$ 
7    $\mathcal{D}_k = \mathcal{U}_k \Sigma_k \mathcal{V}_k^T$  ; // Low rank reconstruction of the image.
8    $P_{\text{Low Rank}} \leftarrow \mathbf{M}(\mathcal{D}_k)$  ; // Classify the reconstructed image.
9    $S(\mathcal{D}) = \text{SSIM}(\mathcal{D}, \mathcal{D}_k)$  ; // Calculate SSIM score.
   /* Detection of adversarial attack */
10  if Prediction Probability  $P_{\text{Original}} = P_{\text{Low Rank}}$  then
11    | No adversarial attack detected.
12  end
13 end
```

3.2 Experimental Setup

All the experiments are performed on Python version 3.10. DNN models trained using Tensorflow and Pytorch platforms. Pytorch attacks are implemented using the Torchattacks 3.5.1 package, and attacks in Tensorflow are generated using coded functions.

DNN Models: Three DNN models are trained for classification on these three datasets. For the Eyepacs dataset, we utilized the ResNet-50 architecture pre-trained on the ImageNet dataset as the backbone. The original top layer of the network is substituted with a new dense layer containing 128 neurons, followed by a dropout layer with a rate of 0.2, and another dense layer with a single neuron for binary classification. This configuration resulted in a validation accuracy of 86.7%. For the IDRID dataset, we use the ResNet-V2 architecture pre-trained on the ImageNet dataset. This is followed by a global average pooling layer with a dropout rate of 0.55. Additionally, a dense layer consisting of 60 neurons and a

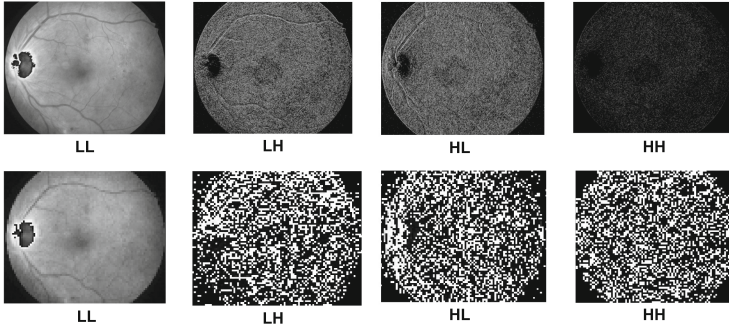


Fig. 4. Image displays the results of a wavelet transform on an original image (top row) and its corresponding adversarial image (bottom row). Each row shows the four decomposed frequency bands.

dropout layer with a rate of 0.3 are added. Finally, a dense layer with 5 neurons and softmax activation is appended for multi-class classification. This model achieved a validation accuracy of 93.6%. For the APTOS dataset, INCEPTION-V3 architecture trained on imagenet is used as the backbone followed by a dense layer of 128 neurons with a drop out of 0.3 and a final layer of two neurons for binary classification. This model achieved a validation accuracy of 94.8%.

Creating adversarial images: Adversarial images are created from the original datasets by employing FGSM and PGD attacks. 2000 adversarial images from the EyePACS dataset, 413 from the IDRID dataset, and 2048 from APTOS are generated using FGSM and PGD attack frameworks. The success rate of attacks for binary and multi-class settings is evaluated using the respective pre-trained models. Results are included in Table 2.

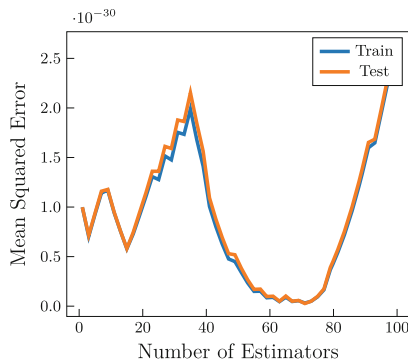


Fig. 5. Plot showing the training and testing errors (MSE) of random forest regressor versus the number of estimators.

Table 2. Results of the proposed method evaluated on the three datasets under multi-class and binary-class classification settings.

Metric	IDRID (Multi-class)		Eyepacs (Binary)		APTOS(Binary)	
	FGSM	PGD	FGSM	PGD	FGSM	PGD
Model Accuracy (%)	93.6	93.6	86.7	86.7	94.8	94.8
Success Rate of Attacks (%)	87.4	89.3	99.6	92.3	92.2	96
Original-Low Rank Match (Clean)	407	407	1682	1682	1940	1940
	413	413	2000	2000	2048	2048
Original-Low Rank Match (Adv.)	42	29	342	242	1721	1820
	361	369	1992	1846	1888	1966
Classification Type	Multi-class	Multi-class	Binary	Binary	Binary	Binary
Attack Detection Accuracy (%)	88.4	92.1	82.8	86.9	91.1	92.57
Average SSIM Score	0.925	0.918	0.88	0.93	0.89	0.92

Entropy Calculation: Actual and adversarial images from both the datasets are converted into YCbCr format followed by wavelet transform of the Y component. Simple Haar wavelet is used to perform the wavelet transform. The entropies of all the four bands are calculated for all the input images and stored for training the regressor for predicting the energy threshold value. Fig. 4 visualizes the wavelet decomposition of the Y component of both the original and adversarial images. The four sub-bands are displayed for each image.

Training the linear regressor: The linear regressor is trained based on the experimentally calculated threshold values and entropies across different datasets in order to generalize the procedure. We built a dataset of 8800 images to train the model. Half the images are real (2000 from Eyepacs, 400 from IDRID and 2000 from APTOS), and the other half are adversarial versions of real images (2000 from Eyepacs, 400 from IDRID, and 2000 from APTOS). It is then separated into training and testing sets to train the model. We noticed that within a specific dataset, changes in entropy result in only minor variations in the threshold value. However, significant differences in threshold values are observed across the datasets. For example, it is observed that the average entropy of the LL band of Y component of the IDRID dataset ranges from 11.25 to 11.57, depending on the attack strength, while that of the Eyepacs dataset falls between 13.12 and 13.26. Consequently, the threshold value ranges from 89% to 91% for the Eyepacs dataset and from 98% to 99% for IDRID. We utilize the entropies obtained from all four bands to train the regressor to make the training more robust.

The Random Forest regressor is employed as the linear model and trained on both actual and adversarial images from the dataset. Subsequently, the Mean Squared Error (MSE) for both the training and testing sets are computed and visualized these metrics against the number of estimators as shown in Fig. 5. Analysis of the plot revealed an optimal number of estimators at approximately 70. This configuration minimizes prediction errors on unseen data, ensuring robust model generalizability.

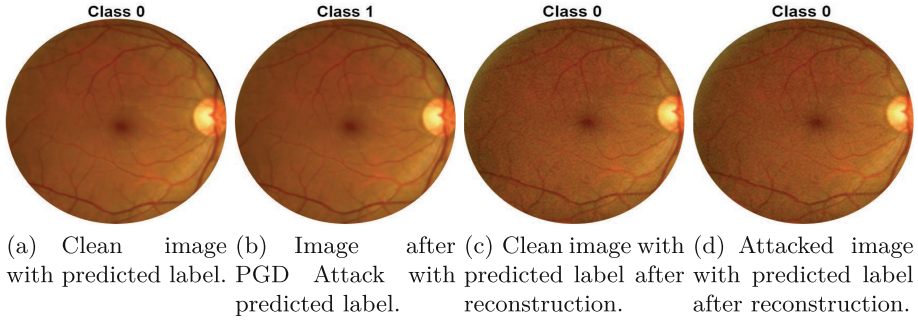


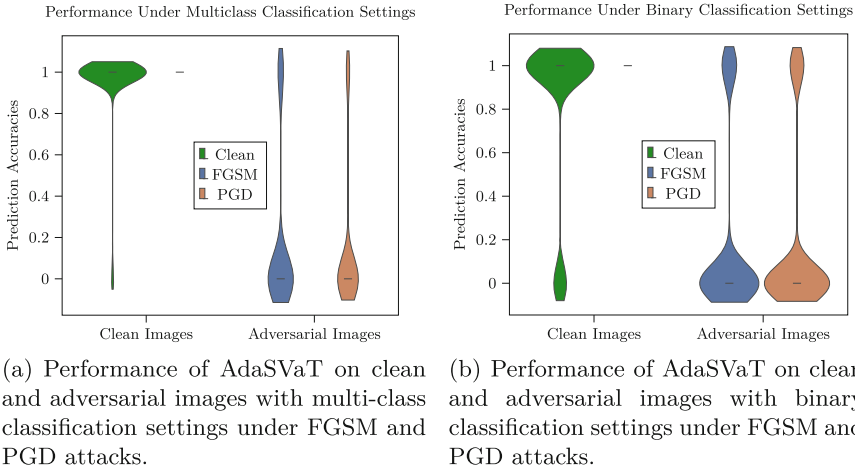
Fig. 6. The resultant images on different stages of AdaSVaT. It can be seen that the original image retained the predicted class label after the reconstruction while the attacked image did not.

SVT and Image Reconstruction: Based on the energy threshold obtained from the linear classifier, hard thresholding of the singular values are performed by retaining only the singular values required to maintain the energy threshold and the image is reconstructed using the thresholded singular values. Resultant images from each stage are shown in Fig. 6. Further, the SSIM score of the input and the output image are calculated in order to ensure that the data loss is not too much during the reconstruction.

3.3 Quantitative and Qualitative Analysis

The evaluation results on the three datasets are presented in Table 2. The proposed method demonstrates its efficacy by achieving over 80% accuracy on all three datasets under two distinct attack types. Attack accuracy is evaluated by removing unsuccessful attacks. The FGSM attack's strength ($\epsilon = 0.03$) is deliberately set higher, as it is relatively less potent compared to PGD. Conversely, the strength of the PGD attack ($\epsilon = 0.003$) is intentionally kept lower. It is noted that the images maintain a healthy similarity score even after the reconstruction. It is also noted that the method performs equally well in both binary and multi-class settings. The spread of accuracies under different attacks for both clean and adversarial images are demonstrated in Fig. 7. The findings reveal that original images retain their predictions post low-rank reconstruction, whereas adversarial images do not. This disparity facilitates the identification of adversarial instances.

The proposed method is benchmarked against the state-of-the-art adversarial attack detection method for fundus images presented at MICCAI 2022 [16], indicating superior performance with minimal computational resources. Unlike existing methods often tailored to specific datasets and classification settings, AdaSVaT demonstrates consistent performance across three fundus image datasets under both binary and multiclass settings (Table 3). It must be noted that all major detection methods operates exclusively under binary classifica-



(a) Performance of AdaSVaT on clean and adversarial images with multi-class classification settings under FGSM and PGD attacks.

(b) Performance of AdaSVaT on clean and adversarial images with binary classification settings under FGSM and PGD attacks.

Fig. 7. Prediction accuracies of clean and adversarial images of both the datasets after low-rank reconstruction under different classification and attack settings.

Table 3. Comparison of AdaSVaT with State-of-the-Art Method. AdaSVaT demonstrates superior accuracy with lower complexity and computational cost.

Criteria	SEViT [16]	AdaSVaT
Mean Detection Accuracy FGSM ($\epsilon = 0.03$)	0.718	0.874
Mean Detection Accuracy PGD ($\epsilon = 0.003$)	0.928	0.905
Number of Datasets Used	One (APTOS)	Three (Kaggle Eyepacs, IDRID & APTOS)
Overall Accuracy	0.823	0.889
Training of Large DNN Models for Detection	✓	✗
Knowledge about the DNN Classifier Model	✓	✗
Multi-class	✗	✓
GPU	✓	✗

tion settings [7, 15, 16]. This highlights the stability of AdaSVaT against diverse classification settings and datasets, contrasting with the state-of-the-art’s fluctuating performance and limited generalizability. Furthermore, AdaSVaT trains a simple linear regressor for attack detection, reducing complexity compared to training a DNN. Additionally, our approach operates independently of the DNN classifier used, unlike existing methods that require partial or complete knowledge. Consequently, AdaSVaT offers simplicity and lower computational requirements.

3.4 Statistical Robustness Tests

We further validate our idea of using lower SVD for detection of the adversarial attacks using statistical tests. We therefore select a random set of 2000 clean images and their corresponding adversarial images. We hypothesize that the distribution of both of these are significantly different in at least 100 lower singular values. We therefore run paired t-tests on singular values for clean and adversarial images. We note that for the distribution significantly differs for lower values with $p < 0.05$ for at least 147 singular values across all 2000 image pairs of clean and adversarial images validating our idea of using singular values for adversarial detection.

3.5 Limitations of Proposed Method

While the current work prioritizes common gradient-based attacks like FGSM and PGD, ensuring broad applicability requires extending the evaluation to encompass other adversarial attack methods. AdaSVaT is likely effective against these techniques as well, particularly those that predominantly affect lower singular values. Additionally, it is required to assess its effectiveness with different medical image modalities like MRI and X-ray to solidify the generalizability.

4 Conclusions and Future Scope

This paper proposes a new method for the detection of adversarial attacks on retinal fundus images. This approach builds on adaptive singular value thresholding based low-rank reconstruction of the image to minimize the effect of adversarial noise. To the best of our knowledge, this is the first work in this domain that utilizes adaptive singular value thresholding and classification inconsistency between original and low-rank images for adversarial detection. Future research should focus on recovering the original image from the adversarial counterpart, minimizing information loss. Our initial efforts in this direction include optimizing the similarity score between the original and low-rank images.

References

1. Basavanthally, A.N., Ganesan, S., Agner, S., Monaco, J.P., Feldman, M.D., Tomaszewski, J.E., Bhanot, G., Madabhushi, A.: Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Trans. Biomed. Eng.* **57**(3), 642–653 (2009)
2. USFDA, <https://www.fda.gov/news-event/press-announcements/fdapermits-marketingartificialintelligencebaseddevicedetectcertaindiabetesrelatedeye>
3. M. Xu, J. Yao, Z. Zhang, R. Li, B. Yang. Learning eeg topographical representation for classification via convolutional neural network, *Pattern Recognit.* 105 ,2020
4. Joseph, N., Ameer, P.M., George, S.N., Raja, K., “Making Domain Specific Adversarial Attacks for Retinal Fundus Images,” in: *Computer Vision and Image Processing*. CVIP, Springer Nature Switzerland, 2024

5. I.J. Goodfellow , J. Shlens , C. Szegedy , Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015
6. K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, X. Lin, Adversarial t-shirt! evading person detectors in a physical world, in: European Conference on Computer Vision, 2020, pp. 665-681
7. Ma, Xingjun and Niu, Yuhao and Gu, Lin and Wang, Yisen and Zhao, Yitian and Bailey, James and Lu, Feng, "Understanding adversarial attacks on deep learning based medical image analysis systems," Pattern Recognition, p. 107332, 2021
8. Zhang, Luxia and Wang, Haibo and Li, Quanzheng and Zhao, Ming-Hu, "Big data and medical research in China," British Medical Journal, vol. 360, 2018
9. IBIS, "<https://www.ibisworld.com/industry-statistics/market-size/health-medical-insurance-united-states>," 2023
10. X. Ren, L. Zhang, D. Wei, D. Shen, and Q. Wang, "Brain mr image segmentation in small dataset with adversarial defense and task reorganization," in Machine Learning in Medical Imaging, 2019
11. M. Watson and N. Al Moubayed, "Attack-agnostic adversarial detection on medical data using explainable machine learning," in International Conference on Pattern Recognition (ICPR), 2021
12. D. Wu, S. Liu, and J. Ban, "Classification of diabetic retinopathy using adversarial training," in IOP Conference Series: Materials Science and Eng. vol. 806, 2020
13. Xu, M., Zhang, T., Li, Z., Liu, M., Zhang, D.: Towards evaluating the robustness of deep diagnostic models by adversarial attack. Med. Image Anal. **69**, 101977 (2021)
14. Lal, S., Rehman, S.U., Shah, J.H., Meraj, T., Rauf, H.T., Abdulkareem, K.H.: Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. Sensors **21**(11), 3922 (2021)
15. Q. Yao, Z. He, and S. K. Zhou, "Medical aegis: Robust adversarial protectors for medical images," 2111.10969, 2021
16. F. Almalik, M. Yaqub, and K. Nandakumar, "Self-ensembling vision transformer (sevit) for robust medical image classification," in MICCAI, 2022, pp. 376-386
17. D. Bharath Kumar, N. Kumar, S. D. Dunston, and V. M. A. Rajam, "Analysis of the impact of white box adversarial attacks in resnet while classifying retinal fundus images," in Computational Intelligence in Data Science, 2022, pp. 162-175
18. O. Daanouni, B. Cherradi, and A. Tmiri, "A novel cnn architecture for robust diabetic retinopathy prediction against adversarial attacks," IEEE Access, 2022
19. L. D. Le, H. Fu, X. Xu, Y. Liu, Y. Xu, J. Du, J. T. Zhou, and R. Goh, "An efficient defending mechanism against image attacking on medical image segmentation models," in MICCAI Workshop, 2022
20. Shi, X., Peng, Y., Chen, Q., Keenan, T., Thavikulwat, A.T., Lee, S., Tang, Y., Chew, E.Y., Lu, Z.: Robust convolutional neural networks against adversarial attacks on medical images. Pattern Recogn. **132**, 108923 (2022)
21. X. Li, D. Zhu, Robust detection of adversarial attacks on medical images, in: International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1154-1158
22. Finlayson, Samuel G and Bowers, John D and Ito, Joichi and Zittrain, Jonathan L and Beam, Andrew L and Kohane, Isaac S, "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287-1289, 2019
23. Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X.: Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction. IEEE Trans. Dependable Secure Comput. **18**(1), 72-85 (2021)

24. Y. Wang, X. Li, L. Yang, J. Ma and H. Li, "ADDITION: Detecting Adversarial Examples With Image-Dependent Noise Reduction," in *IEEE Transactions on Dependable and Secure Computing*, 2023
25. Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial Attacks and Defenses in Deep Learning. *Engineering* **6**, 346–360 (2020)
26. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: *ICLR* (2015)



A New AI System for Precise Grading of HCC Based on Analyzing DW-MRI Radiomics and Alpha-fetoprotein as Liver Cancer Clinical Marker

Abdelrhman Elkhoully¹, Ahmed Alksas², Gehad A. Saleh³,
Mohamed Shehata², Abdelrahman Karawia¹, Mohammed Ghazal⁴,
Sohail Contractor⁵, and Ayman El-Baz²(✉)

¹ Mathematics Department Faculty of Science, Mansoura University, Mansoura, Egypt

² Department of Bioengineering, University of Louisville, Louisville, KY, USA
aselba01@louisville.edu

³ Diagnostic and Interventional Radiology Department Faculty of Medicine, Mansoura University, Mansoura, Egypt

⁴ Electrical, Computer Biomedical Engineering Department, Abu Dhabi University, Abu Dhabi, UAE

⁵ Department of Radiology, University of Louisville, Louisville, KY, USA

Abstract. Hepatocellular carcinoma (HCC), – the main form of liver cancer –, is the second global leading cause of cancer-related mortality. LI-RADS is considered the worldwide non-invasive standard method for imaging interpretation and reporting in patients with HCC eliminating the need for biopsy. However, it might be prone to interpretation subjectivity. Therefore, we develop an objective non-invasive AI-based grading system for HCC for appropriate etiology treatment plans. The developed system integrates potential image-based markers that represent the tumor’s morphology, functionality, and appearance/texture with the associated clinical biomarkers. The study encompasses 117 patients diagnosed with HCC and was divided into three different groups (group 1: benign low-grade (LR 1,2), N = 41; group 2: malignant high-grade (LR 4,5), N = 39; and group 3: malignant not HCC (LR-M), N = 37). Diffusion-weighted magnetic resonance imaging (DWI) was acquired for imaging-based markers identification. The developed grading system pipeline includes: i) estimation of morphological markers using a new parametric spherical harmonic model, ii) estimation of appearance/textural markers using a novel rotation invariant circular binary pattern model, iii) calculation of the functional markers by constructing the representative cumulative distribution functions of the estimated apparent diffusion coefficients, and iv) integrating the aforementioned imaging-based markers with the associated clinical biomarkers, known as Alpha-fetoprotein. The integrated markers were optimized to train and test multiple machine learning (ML) classifiers and a hyper-tuned custom CNN. On a randomly stratified train (80%) test (20%) split scheme, the developed obtained an overall accuracy of 88% in differentiating between

the three groups using the integrated markers along with the CatBoost classifier, surpassing the diagnostic performance of individual marker sets, other ML classifiers, and the CNN as well. The obtained results demonstrate the feasibility of the developed system as a novel tool for non-invasive and objective HCC grading.

Keywords: HCC · AI-Based CAD · Circular Binary Pattern · Spherical Harmonics · ADCs · Combined Markers · CatBoost

1 Introduction

The prevalence of hepatic tumors represents a formidable challenge in healthcare, with early detection and accurate differentiation being pivotal for effective management and improving patient outcomes. Hepatocellular carcinoma (HCC), the predominant form of primary liver cancer, emerges as a leading cause of mortality among cirrhotic patients, marking it as a significant global health burden [1]. Positioned as the fifth most prevalent cancer and the second leading cause of cancer-related deaths worldwide, HCC underscores the critical need for early diagnosis. Notably, the five-year survival rate for HCC can exceed 70% with early-stage detection [2,3]. Uniquely, HCC is the sole malignancy where radiological assessments can suffice for diagnosis, eliminating the need for histopathological confirmation. This diagnosis leverages the specificity of imaging features observed in contrast-enhanced CT or MRI scans, with MRI being favored for its superior soft tissue characterization and absence of ionizing radiation [4,5].

The American College of Radiology (ACR) has developed a Liver Imaging Reporting and Data System (LI-RADS) for standardization in interpreting and reporting hepatic observations within at-risk patients [2,6]. Recognizing the critical role of imaging in HCC management, the ACR implemented the LI-RADS to standardize the interpretation and reporting of hepatic observations. Revised for the last time in 2018, LI-RADS has been an integral part of the American Association for the Study of Liver Diseases (AASLD) guidelines since 2011. Mentioning them indicates the system's importance for clinical practice [1,7]. LI-RADS categorizes liver observations into risk categories depending on enhanced imaging criteria, lesion size, and growth rate. LI-RADS, therefore, helps in the differentiation of HCC from other liver malignancies, differentiation that is actually critical to treatment strategy [4].

Aside from the lesion characterization strength of dynamic contrast-enhanced MRI, it provides a high risk frequency for nephropathy and nephrogenic systemic fibrosis (NSF) with gadolinium-based contrast agents in patients with renal insufficiency [8,9]. This limitation has spurred interest in non-contrast techniques like diffusion-weighted imaging (DWI), a rapid, functional imaging method assessing tissue microcellularity and water molecule motion restriction. DWI's growing application in hepatic lesion characterization highlights its potential in differentiating between benign and malignant hepatic tumors [10,11].

Recent advancements underscore the evolving accuracy in liver tumor diagnosis through imaging innovation [12]. Studies by Wei et al. and Chen et al.

[13,14] have validated the diagnostic efficacy of DWI, demonstrating its utility in distinguishing hepatic metastases from benign lesions and predicting HCC histological grades, respectively, with significant statistical support. Furthermore, Ai et al. [15] strongly emphasized that multiparametric histograms of the intravoxel incoherent motion DWI (IVIM-DWI) model have the highest potential in diagnostic applications and therefore offer a new approach to characterizing phenotypes of liver tumor. Two very promising techniques, machine learning (ML) and deep learning (DL), are integrated with imaging for the diagnosis of HCC. Recent developments include deep convolutional neural networks (CNNs) for classification of tumors and ML classifiers for discriminating HCC from benign lesions. Both innovations showcased improved diagnostic performances and therefore offered great clinical potentials [16–18].

2 Related Work

The exploration of non-invasive diagnostic techniques for hepatic tumors, particularly through DWI and the integration of ML and DL algorithms, constitutes a rapidly evolving area of research. Pivotal studies have laid the groundwork for contemporary diagnostic approaches, highlighting their methodologies, findings, and contributions to the field of hepatic tumor diagnosis.

In a comprehensive meta-analysis study, Wei et al. [13] set out to estimate the diagnostic accuracy of DWI in the differentiation between hepatic metastases and benign focal lesions. A review of 858 cases of hepatic metastasis and 440 cases of benign hepatic lesions from 9 studies points out the high sensitivity and specificity of DWI in lesion differentiation. The study further underscores DWI's potential as a reliable non-contrast diagnostic tool, hence forming a critical ground for more research in the field of imaging-based hepatic tumor diagnosis. Chen et al. [14] reported on the predictive ability of DWI to predict the preoperative histological grade of HCC by incorporating data from 11 studies with a total number of 912 HCC cases. These results are not only a contribution to the diagnostic accuracy but also highlight DWI among the different features for therapeutic planning and assessment of the prognosis in HCC patients. Ai et al. [15] explored the utility of a multiparametric histogram analysis from IVIM-DWI for classifying hepatic tumors, including HCC, hepatic hemangioma (HH), and hepatic cysts (HC). They noted the potential that quantitative imaging markers hold towards improving the diagnostic accuracy for liver tumors and marked a remarkable step towards integrating advanced imaging analytics into clinical diagnostics.

The incorporation of ML and DL in hepatic tumor diagnosis represents a frontier in medical imaging research. Zhen et al. [16] designed a deep CNN model that makes high-quality predictive classifications over non-contrast MRI scans, proving DL to have a very discriminative nature to capture complex patterns of visual semantic relations in order to distinguish among benign primary malignant and metastatic tumors. This study highlights the transformative potential of DL in enhancing diagnostic workflows and patient care strategies. Trivizakis et

al. [17] presented a novel 3-D CNN in order to classify the presence of liver cancer from diffusion-weighted MRI data. Thus, the fact that their approach avoids heavy pre-processing with manual feature extraction attains high accuracy, sensitivity, and specificity in proving the feasibility of DL models being applied for liver tumor diagnosis automation and refinements. Wu et al. [18] developed a ML classification model for the differential diagnosis of HCC and HH based on radiomics features extracted from multi-sequence MRIs. Their work contributes to the growing body of evidence supporting the utility of radiomics in capturing the nuanced characteristics of hepatic lesions, further enabling the development of highly accurate diagnostic tools.

These studies collectively highlight the ongoing advancements in imaging and computational analyses for the diagnosis of hepatic tumors. However, existing studies often do not solely rely on DWI, nor do they fully explore the grading of HCC according to LI-RADS, which is crucial for timely and appropriate treatment. Furthermore, the integration of shape, functional, textural, and clinical markers for enhanced diagnostic performance remains underexplored. To bridge these gaps, we propose a novel computer-aided diagnostic (CAD) system designed to distinguish among various HCC types using a unique integration of shape markers with texture markers, functional markers, and clinical biomarkers utilizing a non-contrast imaging modality, namely; DWI. To our knowledge, this is the inaugural system of its kind that aims to advancing the early differentiation of malignant (LR 4,5) from benign (LR 1,2) HCC tumors and those malignant but not necessarily HCC (LR-M), leveraging the latest in imaging and computational analysis. The main contributions of this study can be summarized as follows:

- Estimating morphological markers using a new parametric spherical harmonic model.
- Estimating appearance/textural markers using a novel rotation invariant circular binary pattern model.
- Calculating the functional markers by constructing the representative cumulative distribution functions (CDFs) of the estimated apparent diffusion coefficients (ADCs).
- Selecting the optimal markers to be combined using Gini-impurity approach.
- Integrating the selected imaging-based markers with the associated clinical biomarkers, known as Alpha-fetoprotein.
- Optimizing the integrated markers to train and test multiple ML classifiers, with the CatBoost surpassing all others.

3 Methods

The model depicted in Figure 1 comprises three primary stages. Initially, preprocessing is conducted on DWI scans to identify tumor lesions, which serve as the region of interest (ROI) across different b-values for each participant. Secondly, a set of three feature types is designed to enable quantitative discrimination among

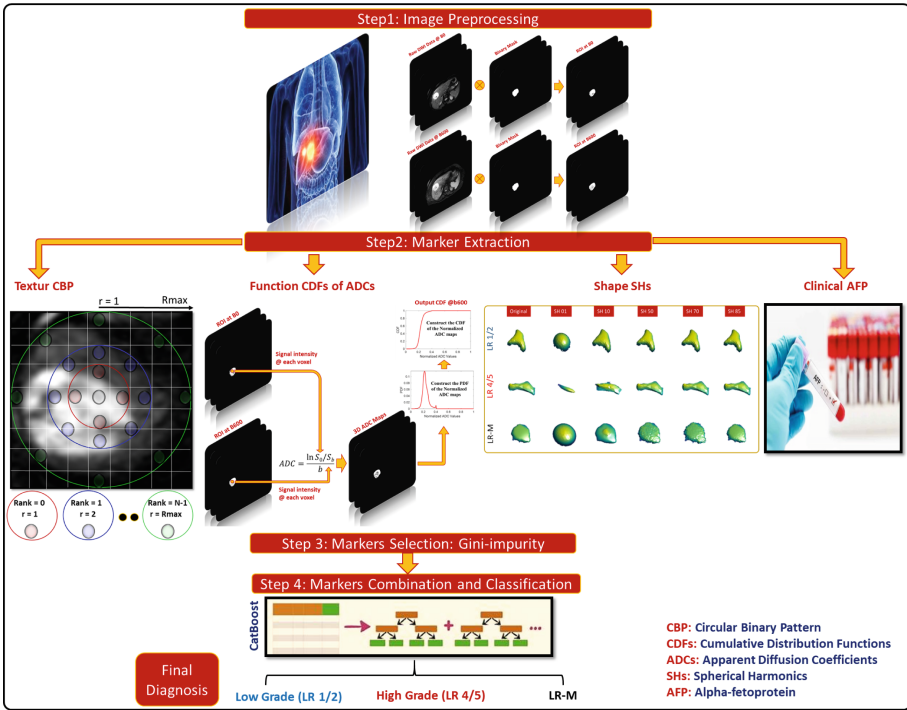


Fig. 1. The proposed pipeline for HCC grading using DWI

the tumor lesions, alongside a clinical biomarker. These include: (i) the circular binary pattern, responsible for estimating appearance/texture; (ii) spherical harmonics-based shape features, describing the shape complexity of liver tumors; and (iii) functional features derived from measuring the rate of water diffusion within tissues at given b-values, known as ADC. Furthermore, these imaging-based markers are then integrated with associated clinical biomarkers, notably Alpha-fetoprotein. Finally, a ML classifier is utilized to combine all the aforementioned individual features to determine the ultimate diagnostic outcome.

3.1 Image Preprocessing

At this stage, the primary objective is to locate tumor lesions and prepare them for the subsequent phase. This involves identifying the ROI for further analysis. For each participant, the dataset initially consists of multiple images acquired at the b-values of 0 and 600 s/mm². The medical team provides definitions of the tumors as binary images, which outline the tumor regions within each DICOM image using our proprietary software. To derive the intended ROIs, these binary masks are applied to the original DICOM gray-scale images. The markers modeled are then extracted and computed from these masked images, resulting in 3D structures representing the tumor lesions for each individual.

3.2 Marker Extraction

Marker extraction is a critical step in the ML workflow, where relevant information is extracted from raw data to create meaningful markers for model training. A good ML marker is a measurable aspect or attribute of a given characteristic that provides valuable information independently. Selecting high-quality markers of objects increases the power of the ML model and makes better decisions. Therefore, the main aim of this step is to transform the preprocessed data into standardized, and machine-understandable markers to distinguish between various subjects and demonstrate to our learning algorithm how to capture the attributes of the tumors. Per consultation with the medical team, these markers are texture/appearance markers, shape markers, functional markers, and clinical biomarkers [19].

Texture Markers: This study introduces Circular Binary Patterns (CBP), a novel texture analysis methodology, for predicting HCC grades from DWI scans. Texture analysis plays a pivotal role in medical image processing [20], offering insights into the structural variations and patterns within tissues or materials. The CBP approach is conceptualized to enhance texture analysis by incorporating rotation invariance and adaptability in identifying texture patterns through Local Binary Patterns (LBP) extension/modification [21]. The fine details of textures are essential to distinguish the different grades of HCC, and these are represented by comparing the intensity of a central pixel with that of its neighbors on concentric circles and thresholding for the differences.

The texture features of the image are evaluated by CBP analyzing their relationship with the surrounding pixels of the present pixel under consideration (x, y) within the concentric circles of radii r from 1 to the maximum radius R_{\max} ; it will be found that all the most significant intensity variations have been correctly identified by the above as valid features of the image texture in a manner independent of its rotation. The updated pixel value, $V_{\text{new}}(x, y)$, is determined as follows:

$$V_{\text{new}}(x, y) = \sum_{r=1}^{R_{\max}} \left(\sum_{(i,j) \in C_r(x,y)} W(i, j, x, y, T) \cdot b^{r-1} \right) \quad (1)$$

where $C_r(x, y)$ denotes the neighbor pixels within radius r , and $W(i, j, x, y, T)$ is a weighting function:

$$W(i, j, x, y, T) = \begin{cases} 1 & \text{if } |V(i, j) - V(x, y)| > T \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Such threshold T is empirically tuned in such a way as to assure the utmost accuracy in discriminating the texture patterns relevant to the different tumors captured by the model. The marker/feature vector for each subject's volume of interest (VOI) is derived from the CDFs or percentiles of the post-CBP transformation. This statistical representation captures the global distribution of texture

markers, providing a robust basis for classifying HCC based on their unique textural characteristics. Algorithm 1, outlines the steps for extracting CBP texture markers from DWI scans and Fig. 2 demonstrates a typical example for the CBP calculation process.

Algorithm 1 Texture Analysis in DWI Scans using Circular Binary Patterns

Require: VOI of DWI scans, maximum radius R_{\max} , threshold T , base b

Ensure: Marker/Feature vector for each VOI

- 1: Calculate circle masks and centers for each pixel
 - 2: **for** each slice in VOI **do**
 - 3: **for** each pixel in slice **do**
 - 4: **if** pixel value > 0 **then**
 - 5: Apply CBP to calculate new pixel value, capturing texture (Equations 1, 2)
 - 6: **end if**
 - 7: **end for**
 - 8: **end for**
 - 9: Construct marker/feature vector from CDFs/percentiles of the transformed VOI, analyzing texture
-

The introduction of CBP for texture analysis in DWI scans represents a significant advancement in medical image analysis. By leveraging the spatial context and intensity information, CBP extracts comprehensive texture markers that are essential for accurately grading HCC.

Functional Markers: ADCs measure the Brownian motion (diffusion) of water molecules within the soft tissues [22]. In malignant tissues that are highly cellular, free water molecules' diffusion is constrained which results in reduced ADCs when compared to benign tumors. In particular, higher grades (e.g., LR4, LR5, and LR-M) are generally related to reduced ADCs, as water diffusion is significantly decreased in the setting of increased tumor cellularity. Relying on this fact, voxel-wise ADCs are computed and ADC maps are constructed from DWI scans at a given b-value using the following equation [23]:

$$\text{ADC} = \frac{\ln(S_0/S_b)}{b} \quad (3)$$

where b refers to the utilized b-value that determines the degree of diffusion weighting applied during the imaging process, which is 600 in our case. S_0 represents the voxel signal intensity at the baseline, b-value equal to zero, and S_b the voxel signal intensity at the non b_0 value, which is 600 in our case. The calculated ADCs are then mapped to their representative CDFs to overcome the challenges that arise due to tumor sizes variability between different subjects and to reduce the training time expenditures, especially in large tumor volumes. The following Fig. 3 is an illustrative example for the ADC maps construction process and its representative CDF at a given b-value.

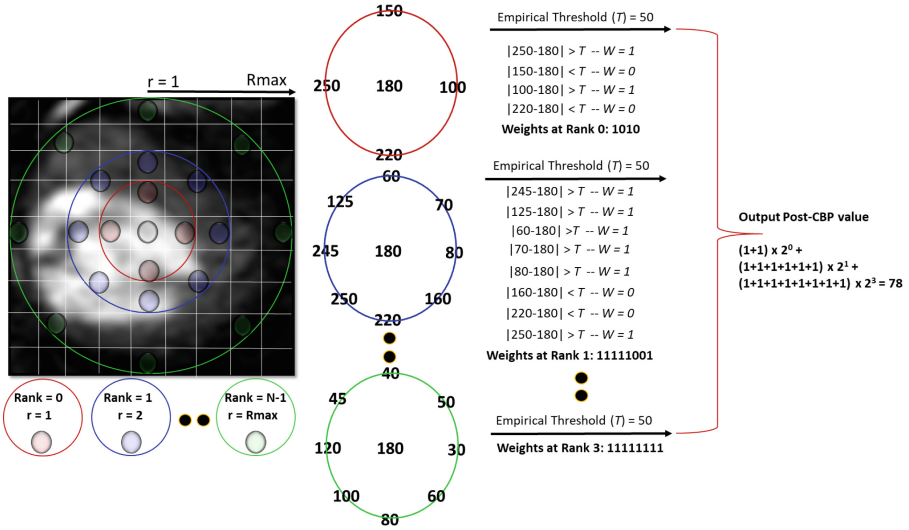


Fig. 2. An illustrative example for the circular binary pattern (CBP) calculations

Shape Markers: The morphological characteristics of HCC tumors exhibit significant variability contingent upon both the degree and the specific type of malignancy, presenting a compelling avenue for enhancing automated diagnostic methodologies through precise shape analysis. In that light, the use of spherical harmonics (SHs) for shape description develops as a powerful mean, providing a sophisticated mathematical framework to capture the intricate geometrical nuances of tumor surfaces.

Spectral spherical harmonics analysis in this case refers to the method of obtaining well-defined shape markers very crucial in the effective detection of liver tumors. This method involves representing the surfaces of the tumor as a sum or linear combination of a series of fundamental functions. Specifically, the initial step involves the construction of a triangulated mesh that meticulously approximates the tumor’s surface topology. Subsequent to this, the mesh is mapped onto a unit sphere utilizing spherical harmonics modeling, a process critically enhanced by our proprietary Attraction-Repulsion Algorithm. This algorithm plays an instrumental role in ensuring the fidelity of the modeling process, by maintaining a uniform distance, typically unitary, between each mesh node and the nodule’s centroid, thereby facilitating a consistent and homogeneous distribution of neighborhood distances among the nodes[24, 25].

Let I denote the totality of mesh nodes, with α symbolizing the cycle counter, and $c_{\alpha,i}$ representing the coordinates of node i at cycle α , where $i = 1, \dots, I$. Furthermore, J embodies the count of neighboring nodes for any given mesh node, and the term $d_{\alpha,ji} = c_{\alpha,j} - c_{\alpha,i}$ delineates the vectorial movement from node j to node i during cycle α . The dynamics of each surface node are governed by

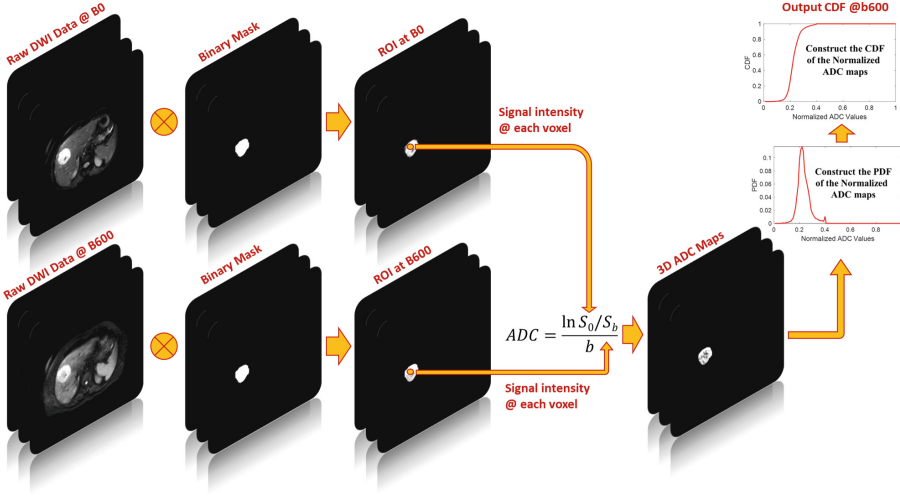


Fig. 3. An illustrative example for voxel-wise ADC (i.e., ADC at each voxel) calculations and the representative cumulative distribution function (CDF) construction

the attraction and repulsion factors, denoted as $c_{A,1}$, $c_{A,2}$, and c_R , which meticulously calibrate the positional adjustments of each node c_i to ensure optimal alignment with its neighbors:

$$C'_{\alpha,i} = C_{\alpha,i} + C_{A,1} \sum_{j=1, j \neq i}^J \mathbf{d}_{\alpha,ji} d_{\alpha,ji}^2 + C_{A,2} \frac{\mathbf{d}_{\alpha,ji}}{d_{\alpha,ji}} \quad (4)$$

Further elucidation on this novel mapping technique is available in [24,25]. Consequent to the mapping endeavor, each tumor’s shape is deciphered through a linear combination of spherical harmonics (with $N = 85$ SHs utilized for this purpose). It is noteworthy that benign tumors, typically manifesting simpler geometrical forms, are characterized using a lower-order combination of spherical harmonics. Conversely, malignant tumors, which are indicative of more complex morphological traits, necessitate a higher-order combination of SHs for accurate representation. This differential approach facilitates a nuanced shape approximation for both benign and malignant tumors, thereby enhancing the diagnostic precision. Please see Fig. 4 that demonstrates the shape complexity differences between individual cases from different groups (i.e., benign (LR1/LR2), malignant HCC (LR4/LR5), and LR-M).

Clinical Biomarkers: According to the National Library of Medicine, the alpha-fetoprotein (AFP) in serum is currently an accessible diagnostic marker for HCC detection. As for patients with chronic liver disease, a sustained increase in AFP serum level was demonstrated to be one of the HCC risk factors and has been used to help identify a high-risk subgroup of chronic liver disease. In

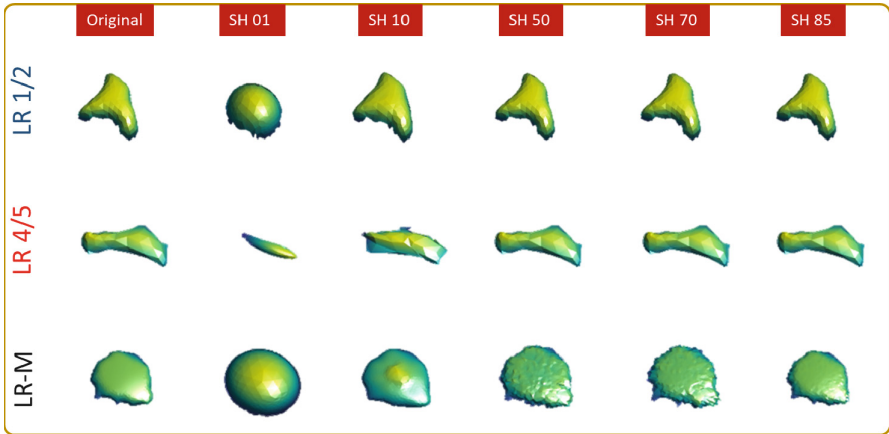


Fig. 4. An illustrative example for spherical harmonics reconstruction process showing shape/surface complexity differences among various HCC groups

patients with liver cirrhosis, fluctuations in AFP levels may reflect the sudden onset of viral hepatitis, the deterioration of the potential liver disease, or the development of HCC [26]. Therefore, the AFP will be included, in our analysis, as a representative clinical biomarker that might be useful for HCC grading.

3.3 Markers selection/importance

Markers selection is a process used to identify the most significant characteristics from a large set of potential markers, 146 in our case. We applied one of the most common methods for markers selection approaches named Gini impurity-based selection, which is completely based on a random forest classifier that is trained to all markers and select the most important/significant ones and return a smaller marker-set [24]. This process ended up with 50 markers, Namely; AFP, CBP ($N = 24$), ADCs ($N = 15$), and SHs ($N = 10$).

3.4 Marker Combination and Liver Tumors Classification

After having the distinct marker sets handy, namely; (i) the circular binary pattern that is responsible for appearance/texture estimation, (ii) the spherical harmonics-based shape markers that can describe the shape complexity of the liver tumors, (iii) the functional markers that are represented by CDFs representation of ADCs, and (iv) and clinical biomarkers, known as Alpha-fetoprotein, now it is time to proceed further with the classification step for distinguishing among three different classes. Namely, benign tumors (LR1/2), malignant tumors (LR4/5), and malignant tumors but not HCC (LR-M). Towards achieving the highest possible diagnostic performance, multiple ML models were optimized for classification tasks (e.g., support vector machine (SVM), random forests (RFs),

naive Bayes classifier (NB), linear regression classifier, CatBoost classifier, and XGBoost classifier). First, classification efficiency was assessed using individual marker sets as shown in Table 1. Then, all of the individual marker sets were combined using concatenation techniques resulting in a combined marker set, which was then used to optimize the same ML classifiers for optimal diagnostic results. It is worth noting that a grid search technique along with the diagnostic accuracy were used as optimization metrics to determine the ideal set of hyper-parameters for various ML classifiers.

Table 1. Enumeration of different marker sets types, names, and number of markers extracted from each type.

Marker Type	Marker Name	Original Markers	Selected Markers
Appearance/Texture	CBP	40	24
Functional	ADCs	20	15
Shape	SHs	85	10
Clinical	AFP	1	1
Total	Combined	146	50

4 Results and Discussion

4.1 Population, MR Data Collection, and Clinical Analysis

Patients with a high risk of developing HCC without a previous history of loco-regional treatment plans were included in this study. All experimental protocols were approved by the University of Louisville, USA, and Mansoura University, Egypt. All participants underwent DWI in the period between August 2021 and May 2023 images, three participants were excluded from the study due to diffusion image quality degradation caused by respiratory motion artifacts. Finally, 117 participants were attained, (M = 55 and F = 62) ranging in age from 36 to 73 years old (average $56 \text{ y} \pm 10 \text{ y}$).

DW images were obtained using a 1.5T Philips Ingenia scanner with a phased-array torso surface coil. All patients were asked to hold their breath during image acquisition to minimize possible respiratory effects. For all participants, DWI was performed using a fat-suppressed single-shot echo-planar sequence with b-values ($b = 0$ and 600 s/mm^2). Diffusion imaging acquisition parameters were as follows: TR/TE = 1900-70 ms, NEX = 3, matrix = 124 120, slice thickness = 5 mm, slice gap = 1-2 mm, and scan time = 70 sec.

According to LI-RADS v2018(4), LI-RADS classified the hepatic observations into LR-1 (definitively benign), LR-2 (probably benign), LR-3 (indeterminate HCC), LR-4 (probably HCC), LR-5 (definitively HCC), and LR-M (malignant but not definitely HCC). In this study, a 15-year hands-on-experience radiologist

performed MR clinical analysis following the aforementioned LI-RADS v2018(4) guidelines to provide the ground truth diagnosis for the participants. Among the 117 participating patients, 41 liver tumors were diagnosed as benign tumors (LR1 = 14 and LR2 = 27), 39 were diagnosed as malignant tumors (LR4 = 19 and LR5 = 20), and 37 were diagnosed as malignant but not necessarily HCC (LR-M) tumors.

4.2 The Proposed CAD Evaluation

The developed CAD system for liver tumors grading was assessed using a randomly stratified train test split approach on the 117 liver tumor subjects. In which 80% were used for training and 20% were saved for testing. All the obtained results were reported and tabulated using the following metrics: the overall accuracy, precision, recall, F1-score, and, weighted F1-score for finding the actual insights of predicted instances of each class. Various combinations of marker sets (see Table 1) were used along with many ML classifiers to find the optimal diagnostic results among all of them. With 88.0% overall accuracy, the CatBoost optimized classification model along with the combined markers set outperformed the performance of all individual marker sets in distinguishing among different classes (LR 1,2), (LR 4,5), and (LR-M) as documented in Table 2. These obtained results demonstrate the advantage of combining all individual marker sets that accurately characterizes the tumor status. This can be justified in part by that the combination process allowed for accounting for different aspects, namely; shape, texture, and function, with the aid of clinical biomarkers as well.

It is important to note that CatBoost classifiers [27], [28] are well-known, reliable ML classification techniques that are frequently applied for solving classification problems, especially in the medical domain [29]. CatBoost is one of the family of GBDT ML ensemble approaches [27] that relies on a gradient-boosting algorithm which is particularly effective for handling data sets with categorical features. In particular, GBDT algorithms enhance the ensemble by incorporating decision trees. They choose the tree that minimizes the loss function $L(\hat{y}, y)$, which operates on the output values of the ensemble \hat{y} and the corresponding labels y . This process occurs during training using a specific dataset. After training, the ensembles output probability is the sum of probabilities associated with the classes indicated in the leaf nodes of the decision trees. For these reasons, CatBoost outperformed all other ML classification models in terms of accuracy, precision, recall, F1-score, and weighted F1-score measures and thus, was selected as the optimal classifier for the intended liver tumor grading problem. A favorable comparison that highlights the superior performance of the CatBoost over different well-known ML classifiers and over a custom-designed CNN as well is shown in Table 3. It is worth noting that the utilized CNN architecture, manually constructed and tailored to the specific needs of the task, is appropriate for the resized medical data consisting of 64x64x32 volumes where each volume is composed of 32 grayscale images.

Table 2. Comparison between classification reports for each category of individual marker sets and the combination of all together

Features	Classes	Precision	Recall	F1-score	Accuracy	Weighted F1-score
ADCs	(LR 1,2)	0.45	0.62	0.53	0.5	0.48
	(LR 4,5)	0.5	0.25	0.33		
	(LR M)	0.56	0.62	0.59		
AFP	(LR 1,2)	0.62	0.73	0.67	0.67	0.66
	(LR 4,5)	0.9	0.9	0.9		
	(LR M)	0.43	0.33	0.38		
SHs	(LR 1,2)	0.67	0.5	0.57	0.54	0.54
	(LR 4,5)	0.45	0.62	0.53		
	(LR M)	0.57	0.5	0.53		
CBP	(LR 1,2)	0.57	0.73	0.64	0.6	0.6
	(LR 4,5)	0.56	0.5	0.53		
	(LR M)	0.71	0.56	0.63		
Combined	(LR 1,2)	1	0.88	0.93	0.88	0.88
	(LR 4,5)	0.88	0.88	0.88		
	(LR M)	0.78	0.88	0.82		

Table 3. Comparison between classification reports for each ML classifier that highlights the advantage of the utilized CatBoost classifier over others

ML Model	Classes	Precision	Recall	F1-score	Accuracy	Weighted F1-score
SVM	(LR 1,2)	0.78	0.88	0.82	0.79	0.79
	(LR 4,5)	0.86	0.75	0.80		
	(LR M)	0.75	0.75	0.75		
Random Forest	(LR 1,2)	0.86	0.75	0.80	0.75	0.75
	(LR 4,5)	0.67	0.75	0.71		
	(LR M)	0.75	0.75	0.75		
XGBoost	(LR 1,2)	0.67	1	0.80	0.75	0.72
	(LR 4,5)	0.78	0.88	0.82		
	(LR M)	1	0.38	0.55		
CNN	(LR 1,2)	0.78	0.88	0.82	0.79	0.8
	(LR 4,5)	1	0.75	0.86		
	(LR M)	0.67	.75	0.71		
CatBoost	(LR 1,2)	1	0.88	0.93	0.88	0.88
	(LR 4,5)	0.88	0.88	0.88		
	(LR M)	0.78	0.88	0.82		

5 Conclusions, Limitations, and Future Work

In summary, the proposed AI-based CAD system demonstrated feasibility and efficacy for HCC grading with an acceptable overall accuracy of 88.0% on a train

test split criteria. It has shown promising results given that a non-contrast MR imaging modality (DWI) was used. The developed AI-based CAD system relies on the integration of different types of individual marker sets that account for multiple aspects to fully characterize the tumor including shape, texture, function, and clinical descriptors for precised quantification and diagnosis. Clearly, the final diagnostic results were much improved by the meaning of markers integration rather than depending on individual marker sets. In addition, the Cat-Boost ML classification model demonstrated the optimal classification ability among all machine learning classifiers, being recommended for such multi-class classification problems. This study was limited by not including LR3 group due to the lack of collection of this rare class. Moreover, this CAD system is not fully automated in its current form as it still need an expert radiologist to perform manual segmentation of liver tumors from DWI scans and still depend on hand-crafted features. A larger data cohort including LR3 group is currently being collected for further investigation of the extended diagnostic capabilities of the developed CAD system.

References

1. Gehad A Saleh, Ali H Elmokadem, Ahmed Abdel Razek, Ahmed El-Morsy, Omar Hamdy, Elshimaa S Eleraky, and Marwa Saleh. Utility of diffusion tensor imaging in differentiating benign from malignant hepatic focal lesions. *European Radiology*, 33(2):1400–1411, 2023
2. Nobuhiro Tsuchiya, Yu., Sawada, I.E., Saito, K., Uemura, Y., Nakatsura, T.: Biomarkers for the early diagnosis of hepatocellular carcinoma. *World J Gastroenterol: WJG* **21**(37), 10573 (2015)
3. Ahmed Abdel Khalek Abdel Razek, Lamiaa Galal El-Serougy, Gehad Ahmad Saleh, Walaa Shabana, and Rihame Abd El-wahab. Liver imaging reporting and data system version 2018: what radiologists need to know. *Journal of Computer Assisted Tomography*, 44(2):168–177, 2020
4. Ahmed Abdel Khalek Abdel Razek, Lamiaa Galal El-Serougy, Gehad Ahmad Saleh, Rihame Abd El-Wahab, and Walaa Shabana. Interobserver agreement of magnetic resonance imaging of liver imaging reporting and data system version 2018. *Journal of Computer Assisted Tomography*, 44(1):118–123, 2020
5. Julie Y An, Miguel A Peña, Guilherme M Cunha, Michael T Booker, Bachir Taouli, Takeshi Yokoo, Claude B Sirlin, and Kathryn J Fowler. Abbreviated mri for hepatocellular carcinoma screening and surveillance. *Radiographics*, 40(7):1916–1931, 2020
6. Khaled M Elsayes, Kathryn J Fowler, Victoria Chernyak, Mohab M Elmohr, Ania Z Kielar, Elizabeth Hecht, Mustafa R Bashir, Alessandro Furlan, and Claude B Sirlin. User and system pitfalls in liver imaging with li-rads. *Journal of Magnetic Resonance Imaging*, 50(6):1673–1686, 2019
7. A-Hong Ren, Peng-Fei Zhao, Da-Wei Yang, Jing-Bo Du, Zhen-Chang Wang, and Zheng-Han Yang. Diagnostic performance of mr for hepatocellular carcinoma based on li-rads v2018, compared with v2017. *Journal of Magnetic Resonance Imaging*, 50(3):746–755, 2019
8. Ledneva, E., Karie, S., Launay-Vacher, V., Janus, N., Deray, G.: Renal safety of gadolinium-based contrast media in patients with chronic renal insufficiency. *Radiology* **250**(3), 618–628 (2009)

9. Stephanie Fox-Rawlings and Diana Zuckerman. Nchr report: the health risks of mris with gadolinium-based contrast agents. *National Center for Health Research Q*, 9, 2020
10. Gehad Ahmad Saleh, Ahmed Abdel Khalek Abdel Razek, Lamiaa Galal El-Serougy, Walaa Shabana, and Rihame Abd El-Wahab. The value of the apparent diffusion coefficient value in the liver imaging reporting and data system (li-rads) version 2018. *Polish Journal of Radiology*, 87:e43, 2022
11. Taron, J., Johannink, J., Bitzer, M., Nikolaou, K., Notohamprodjo, M., Hoffmann, R.: Added value of diffusion-weighted imaging in hepatic tumors and its impact on patient management. *Cancer Imaging* **18**, 1–7 (2018)
12. Arya Haj-Mirzaian, Ana Kadivar, Ihab R Kamel, and Atif Zaheer. Updates on imaging of liver tumors. *Current oncology reports*, 22:1–10, 2020
13. Chenggang Wei, Jieying Tan, Li Xu, Liu Juan, Si Wei Zhang, Lu Wang, and Qun Wang. Differential diagnosis between hepatic metastases and benign focal lesions using dwi with parallel acquisition technique: a meta-analysis. *Tumor Biology*, 36:983–990, 2015
14. Chen, J., Mingpeng, W., Liu, R., Li, S., Gao, R., Song, B.: Preoperative evaluation of the histological grade of hepatocellular carcinoma with diffusion-weighted imaging: a meta-analysis. *PLoS ONE* **10**(2), e0117661 (2015)
15. Zhu Ai, Qijia Han, Zhiwei Huang, Jiayan Wu, and Zhiming Xiang. The value of multiparametric histogram features based on intravoxel incoherent motion diffusion-weighted imaging (ivim-dwi) for the differential diagnosis of liver lesions. *Annals of Transnational Medicine*, 8(18), 2020
16. Shihui Zhen, Weizhi Luo, Zhiyu Jiang, Yankai Jiang, Jihong Sun, Liqing Zhang, Yujun Wang, Zhongyu Wu, Yubo Tao, Ming Cheng, et al. Deep learning-assisted diagnosis of liver tumors using non-contrast magnetic resonance imaging: A multi-center study
17. Eleftherios Trivizakis, Georgios C Manikis, Katerina Nikiforaki, Konstantinos Drevelegas, Manos Constantinides, Antonios Drevelegas, and Kostas Marias. Extending 2-d convolutional neural networks to 3-d for advancing deep learning cancer classification with application to mri liver tumor differentiation. *IEEE journal of biomedical and health informatics*, 23(3):923–930, 2018
18. Jingjun, W., Liu, A., Cui, J., Chen, A., Song, Q., Xie, L.: Radiomics-based classification of hepatocellular carcinoma and hepatic haemangioma on precontrast magnetic resonance images. *BMC Med. Imaging* **19**(1), 1–11 (2019)
19. Guido Van Rossum and Fred L Drake. Python library reference, 1995
20. Scalco, E., Rizzo, G.: Texture analysis of medical images for radiotherapy applications. *Br. J. Radiol.* **90**(1070), 20160642 (2017)
21. Timo Ojala, Matti Pietikainen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585. IEEE, 1994
22. Geetha Soujanya Chilla, Cher Heng Tan, Chenjie Xu, and Chueh Loo Poh. Diffusion weighted magnetic resonance imaging and its recent trend-a survey. *Quantitative imaging in medicine and surgery*, 5(3):407, 2015
23. Le Bihan, D., Breton, E.: Imagerie de diffusion in-vivo par résonance magnétique nucléaire. *Comptes-Rendus de l'Académie des Sciences* **93**(5), 27–34 (1985)
24. Ahmed Alksas, Mohamed Shehata, Gehad A Saleh, Ahmed Shaffie, Ahmed Soliman, Mohammed Ghazal, Adel Khelifi, Hadil Abu Khalifeh, Ahmed Abdel Razek, Guruprasad A Giridharan, et al. A novel computer-aided diagnostic system for accurate detection and grading of liver tumors. *Scientific reports*, 11(1):13148, 2021

25. Williams, E., El-Baz, A., Nitzken, M., Switala, A., Casanova, M.: Spherical harmonic analysis of cortical complexity in autism and dyslexia. *Transl. Neurosci.* **3**(1), 36–40 (2012)
26. Jiaxin Zhang, Guang Chen, Peng Zhang, Jiaying Zhang, Xiaoke Li, Da'nian Gan, Xu Cao, Mei Han, Hongbo Du, and Yong'an Ye. The threshold of alpha-fetoprotein (afp) for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis. *PLoS One*, 15(2):e0228857, 2020
27. Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018
28. John T Hancock and Taghi M Khoshgoftaar. Catboost for big data: an interdisciplinary review. *Journal of big data*, 7(1):94, 2020
29. Dhananjay, B., Sivaraman, J.: Analysis and classification of heart rate using catboost feature ranking model. *Biomed. Signal Process. Control* **68**, 102610 (2021)



Detection of Extremely Sparse Key Instances in Whole Slide Cytology Images via Self-supervised One-class Representation Learning

Swarnadip Chatterjee^(✉), Orcun Göksel, Nataša Sladoje, and Joakim Lindblad

Department of Information Technology, Uppsala University, Uppsala, Sweden
swarnadip.chatterjee@it.uu.se

Abstract. Whole slide pathological image classification using slide-level labels often relies on multiple instance learning. Multiple instance learning based approaches are particularly challenging with whole slide cytology images, where the vast number of instances can make it difficult to identify key instances, especially when they are scarce. In this work we evaluate whether using representations learnt from patches from only normal slides is effective for instance-level decision making. We aim for interpretable slide-level decision making for whole slide cytology images. We focus on the effectiveness of a self-supervised contrastive learning framework within a one-class classifier setting, assessing its ability to learn the appearances of normal cells from a limited number of normal slides and subsequently identify abnormal cells (key instances) on test slides. We evaluate our approach on a publicly available cytology dataset, achieving a *Recall@400* score of **0.1938**, considerably improving over the **0.1109** score obtained using a weakly supervised approach.

Keywords: One-class classification · Contrastive self-supervised learning · Outlier detection · Cytology

1 Introduction

Detection of potentially malignant disorders in the oral cavity at an early stage is critical for successful treatment and improved survival [5, 6, 8, 25]. The high difficulty of the pathological analysis, fatigue and a need for a second opinion when diagnosing diseases, accompanied by variations in pathology, creates opportunities for computer assistance to aid researchers and pathologists. This is driving the development of computational pathology methods that leverage automated image analysis techniques, and in particular deep learning-based approaches – spurred by the recent rapid advancement of the field, to support the analysis.

The widespread adoption of digital slide scanners for scanning histology slides has facilitated a rapid rise in the use of multi-gigapixel whole-slide images (WSIs) (which can be as large as $120,000 \times 120,000$ pixels and may contain as many

as 100,000 cells) for early-stage cancer detection in the field of computational pathology [4]. Whole slide imaging has also gained popularity among cytologists due to its potential to improve diagnostic accuracy and workflow efficiency via analysis of whole-slide cytology images with automated diagnostic systems [1]. Although advanced-stage cancer cytology samples may contain around 30% malignant cells in whole slide images, whole-slide cytology images corresponding to early-stage cancer samples may contain as few as 1% of the cells as malignant. Due to the large number of cells present and the sparsity of malignant cells in early-stage cancers, obtaining cell-level annotations is difficult, time-consuming, and unreliable as they might suffer from inter-observer variability.

The vast majority of methods for automated analysis of WSIs first decompose the very large images into much smaller-sized tiles containing relevant visual information (e.g., nuclei, cells, tissue structures, etc.) and then perform further patch-based processing of the entire slide. Since obtaining patch-based labels is difficult and costly and may also suffer from annotation bias, recent methods are mainly based on deep multiple instance learning (MIL) (often using vision transformers (ViT) and contrastive self-supervised pre-training) for detecting abnormal cells or classifying WSIs [9, 13, 21], as they enable learning based on weakly (per-slide) labeled data while still providing some level of instance-level interpretability.

Multiple Instance Learning (MIL) [15] is a machine learning paradigm where training data is organized into bags, each containing multiple instances. In MIL, a bag is labeled positively if at least one instance in the bag is positive, and negatively otherwise. The goal is to learn a classifier from these bags rather than from individual instances. This approach is particularly useful in situations where only the collective information of a bag determines its label, and the specific positive instances within the bag are unknown or ambiguous. However, MIL-based methods suffer from memory constraints, particularly for very large bags [12].

Self-supervised methods are able to learn generalizable and domain-invariant representations both on a patch-level and slide-level; however, using those representations for downstream classification tasks also requires some amount of patch-based labels. Obtaining such labels for cytology WSIs may be infeasible, as it is time-consuming, costly, and also may vary between observers.

In this study, we evaluate the effectiveness of representations learned from patches belonging only to WSIs corresponding to healthy patients for the task of detecting malignant patches (or cells) on unseen WSIs. Our main assumption is that all the patches belonging to normal slides are normal, while abnormal slides contain an unknown mixture of normal and abnormal patches. We use images of normal cells to learn a representation of the normal class and to develop an abnormal cell detector to identify abnormal patches, aiming for interpretable decision-making for the whole slide (as instance- (or patch-) level decisions can be obtained which can be aggregated to have a slide-level decision). In particular, we evaluate self-supervised representation learning in the one-class setting

since, unlike the binary-class scenario, it does not require any labeled data for performing downstream classification tasks.

The code is available publicly at https://github.com/MIDA-group/occ_cyto.

2 Background and Related Work

Detection of abnormal cells, reliably and efficiently, from whole-slide cytology images is essential for cost-efficient and non-invasive screening for oral cancer. In this section, we briefly review methods proposed for detecting abnormal cells in whole-slide cytology/histology data.

Weakly supervised methods are a common choice for the detection of abnormal patches in whole-slide images. These methods are applicable to data with both slide-level annotations and limited patch-level annotations. Interpretability is another desired aspect of these methods. Attention-Based deep Multiple Instance Learning (ABMIL) [9] is one such method that highlights patches with maximum contributions to the decision-making at the slide level. Other proposed MIL-based approaches involving Self-Supervised Learning and Vision Transformers (ViT) include Dual Stream MIL (DS-MIL) [13], Clustering-Constrained Attention Multiple-Instance-Learning (CLAM) [14], and TransMIL: Transformer-Based Correlated Multiple Instance Learning [21].

MIL-based methods, in general, suffer from memory limitations when applied to whole-slide cytology images, which contain a large number of instances [12]. The large memory requirement of ABMIL can be circumvented either by sampling or by gradient accumulation [2]. Although being data-hungry, ViT-based methods, similar to MIL-based ones, are able to successfully aggregate patch-based features to form good slide-level representations. However, unlike whole-slide histology images, whole-slide cytology images (where the patches contain at most a few overlapping cells) do not have much contextual information (and detection of a few key instances present is the main goal), which might be the reason for fewer applications of ViTs on cytology data, as ViTs are known to be mostly useful in aggregating contextual information, which makes them more suitable for histology whole-slide images. In a recent study [11], it is shown that ABMIL does not work well for a small ratio of key (i.e., positive/abnormal) to normal instances in large bags. It compares Single Instance Learning (SIL), which involves training on individual instances with specific labels (normal/malignant), with Multiple Instance Learning (MIL), which uses bags of instances, assigning labels based on the presence of at least one positive instance in the bag. Although SIL performs better than ABMIL (which also suffers from mode collapse) on oral cancer data, both SIL and ABMIL struggle when the number of key instances is less than 2% in large bags.

We explore if a One-Class Classification (OCC) approach can be used to overcome the limitation of struggling to identify key instances when their percentage (also known as “witness rate”) is less than 2%. OCC is an approach for abnormal instance detection, which involves fitting a model on a single class (normal cells) and then predicting whether a test data point is from the normal cells or

not. One of the most popular deep architectures used for this is Deep One-Class Classification [18], which is trained on an anomaly detection-based objective. However, this approach usually tends to map all the normal data into a single point-known as "hypersphere collapse" [3]. To avoid hypersphere collapse and learn effective one-class representations, [23] propose a distribution-augmented contrastive learning-based two-stage framework for building deep one-class classifiers.

Another approach for outlier detection is to use deep generative models [17, 27, 28]. However, for these methods, accurate density estimation is problematic in high dimensions [24]. Moreover, learning deep generative models on raw image data is difficult, as they tend to assign high importance to background pixels and also learn local pixel correlations [10]. To make these methods work, a good representation of the image data is needed first.

3 Data

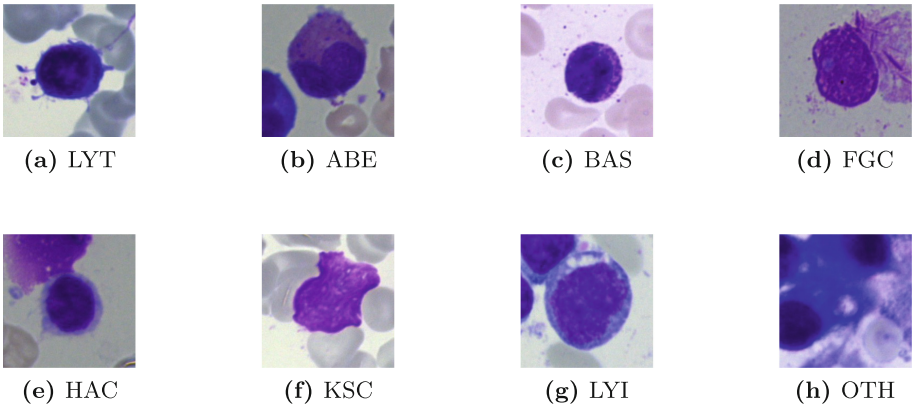
To enable a well-controlled evaluation of methods, while still working with realistic data, we create a synthetic dataset consisting of real cells. We conduct experiments on a subset of the publicly available PKGBM bone marrow cytology dataset [16] (containing cells belonging to 21 categories). We choose this particular dataset as it contains cell sized patches divided into well defined classes and the number of instances in some of the majority classes approach the number of cells present in real whole slide cytology images. To recreate a realistic WSI-like scenario, we choose a particular majority class from all the available classes to represent the normal cells in a whole-slide cytology image, and pick all the minority classes as abnormal cells. Particularly, we consider the *Lymphocyte (LYT)* category as *normal* and the following 7 classes as abnormal: *Abnormal eosinophil (ABE)*, *Basophil (BAS)*, *Faggott cell (FGC)*, *Hairy cell (HAC)*, *Smudge cell (KSC)*, *Immature lymphocyte (LYI)*, and *Other cell (OTH)*. Moreover, the reason we choose LYT as the normal class is that both LYT and LYI are present in the dataset, and they can represent real-life whole-slide cytology images having normal and abnormal cells belonging to a particular cell type. The other six minority classes are included as abnormalities because, in a whole-slide cytology image, there are other abnormalities present that are not of the same cell type. The names of the categories of cells and their numbers are summarized in Table 1. All the images from each of the categories are of size 250×250 pixels. Sample images (one from each cell type) are presented in Figure 1.

4 Method

We adapt the approach suggested by [23] for contrastive self-supervised OCC to identifying abnormal cells in cytology images, where an encoder is trained using cell-sized patches from only normal whole-slide cytology images. We identify suitable augmentation approaches for augmenting the distribution of the training

Table 1. Summary of the categories of our dataset

Cell Type	Total	Train	Test
LYT (normal)	26242	18369	7873
BAS (abnormal)	441	308	133
HAC (abnormal)	409	286	123
OTH (abnormal)	294	205	89
LYI (abnormal)	65	45	20
FGC (abnormal)	47	32	15
KSC (abnormal)	42	29	13
ABE (abnormal)	8	5	3

**Fig. 1.** Sample images of the different cell types from the PKGBM dataset

data and pre-training the encoder. We evaluate four different *strong* augmentations for augmenting the distribution of the training data, in order to reduce the uniformity of the learned normal class representations. Also, for training the encoder with the distribution-augmented normal cell data, we use relevant *weak* augmentations, namely `RGBShift` (to handle stain variations across cytology slides) and standard `RandomResizedCrop` (to handle small variations in the size of the nuclei and cells due to varying levels of fluid absorption during the cytology slide preparation process). The learned representations of the normal cells are then used to train a One-Class Support Vector Machine (OC-SVM) for one-class classification of normal cells. We elaborate on the steps of our method below.

4.1 Self-supervised representation learning

For self-supervised pre-training using only normal cell patches, we adapt the recently proposed distribution-augmented contrastive learning method that

extends training data distributions via data augmentation [23]. This has been proven to be particularly effective in learning representations for OCC, as it reduces the class collision between examples from the same class and the uniformity of the learned embeddings on the unit hyper-sphere.

Contrastive learning learns representations by distinguishing different views (e.g., augmentations) of the same instance from other data instances. Two different views of the same object are often referred to as a positive pair. The contrastive loss for a positive pair $\{p_i, p_j\}$ is given by

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{p}_i, \mathbf{p}_j)/\tau)}{\sum_{\forall k \neq i} \exp(\text{sim}(\mathbf{p}_i, \mathbf{p}_k)/\tau)}, \quad (1)$$

where

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \quad (2)$$

is the *Cosine Similarity* between two vectors \mathbf{x} and \mathbf{y} , and τ is the temperature coefficient. The final loss is computed across all positive pairs, both (i, j) and (j, i) , in a mini-batch.

While self-supervised representation learning using a contrastive loss as above has proven effective for multi-class classification, its ineffectiveness for one-class classification, due to class-collision and uniformity [19, 23, 26], prompts the modification of traditional contrastive learning frameworks for one-class classification via:

1. The use of *moderate* batch size;
2. Performing *distribution augmentation* by incorporating *weak* and *strong* augmentations in order to address the class collision and uniformity issues. The weak augmentations help in the self-supervised learning process while the strong ones help augment the distribution of the normal class data.

For our one-class (i.e. normal class) classification problem we explore the following augmentations for cell patches:

Weak augmentations:

```

RGBShift: (r_shift_limit=10, g_shift_limit=10,
           b_shift_limit=10, p=1),
RandomResizedCrop: (224, 224, scale=(0.9, 1.0), ratio=(1, 1),
                    interpolation=cv2.INTER_LANCZOS4, p=1).
    
```

Strong augmentations:

```

CenterCropResize: CenterCrop(height=180, width=180, p=1),
                  Resize(height=250, width=250, p=1).
ColorJitter: (p=1).
GridDistortion: (p=1).
ElasticTransform: (p=1).
    
```

With this **weak and strong** augmentation scheme, aimed at making the inlier distribution become more compact [23] (as it pushes the normal class representations closer with the help of the strong augmentation generated additional training data), we obtain representations learned in a self-supervised manner (using only the normal samples).

4.2 One class classifier with learned representations

To construct our final abnormal cell detector (using a one class classifier learned using normal cells), we train a One Class SVM (OC-SVM) [20] using the learned representations. For this, we choose the *Radial Basis Function (RBF)* kernel. This kernel involves hyper-parameters ν , that sets an upper bound on the fraction of training errors (thus setting a lower bound on the fraction of support vectors), and also the RBF kernel coefficient γ , which defines the kernel width.

For comparing the performance of both the fully supervised and weakly supervised methods, we use the same train and test sets as used in the contrastive self-supervised one-class learning setup.

4.3 Fully supervised

We use ResNet18 [7] as a base network in all experiments performed. Based on the performance in the supervised experiment (Sect. 6), we judge it to be an architecture strong enough for the task.

In order to compare the proposed approach for outlier detection in cytology with a fully supervised method (since cell level annotations are present for the dataset used for this work), we train a ResNet18 on the cytology dataset with a 70-30 train-test split. For a fair comparison, we use the same weak augmentations i.e. `RGBShift` and `RandomResizedCrop` for the training set.

4.4 Weakly supervised setup

We also conduct experiments by simulating a weakly labelled scenario by bagging normal cells with abnormal ones for some bags and keeping only normal cells in the others. We label a bag as `abnormal` if it contains one or more abnormal cells and assign the same label as that of the bag to all the instances of that bag. We then train a ResNet18 using the individual instances of each bag.

5 Experiments

5.1 Experimental setup

We perform three experiments for the classification of cells: one fully supervised, one weakly labelled, and one one-class trained. The fully supervised learning experiment is included to provide a reference for what is an expected upper bound of performance in case we actually have access to instance level labels. The weakly labelled experiment relies on bag level labels and corresponds to the method presented in [11]. The one-class trained approach is what we propose in this paper as an alternative to the weakly labelled approach.

In the fully supervised setup, we train a pre-trained (on ImageNet) ResNet18 on our dataset. We use augmentations `RGBShift` and `RandomResizedCrop` for the training set. `RGBShift` modifies the intensity of each channel of the RGB image by ± 10 , whereas `RandomResizedCrop` modifies it by a scale of 0.9 to

1.0 using (Lanczos interpolation with a window size 4). These augmentations are applied to all instances in the training set (with probability $p = 1$). We train seven different binary classification models, each corresponding to one of the seven abnormal classes (or outlier sets), using cross-entropy loss and the Stochastic Gradient Descent (SGD) optimization algorithm. To evaluate the performance of the OCC framework for different types of outliers, one at a time, we train seven different binary classifiers, with each scenario having only one abnormality present in the dataset. Similarly, in the bagged learning setup, we also trained seven different models, each with one particular abnormal class.

Table 2. Summary of the percentage of key instances in the weakly supervised setup

Abnormal Class	Number of key/bag	Percentage of key
BAS	62	3.353%
HAC	57	3.113%
OTH	41	2.232%
LYI	9	0.489%
FGC	6	0.348%
KSC	6	0.315%
ABE	1	0.054%

In the bagged learning setup, the total 18369 normal (LYT) images are divided into 10 approximately equal bags. Five of those are mixed with equal number of random samples drawn from a single outlier class. We label a bag as *abnormal* if at least one sample from the outlier set is present and *normal* otherwise. We train a ResNet18 instance-wise (using bag labels) using the same augmentations and other settings as for the fully supervised one and test it instance-wise on the test set. The percentage of key instances present in the created positive bags are presented in Table 2.

In the self-supervised scenario, we use the *strong* augmentations as described in Section 4.1 for augmenting the training data distribution. For training the ResNet18 encoder using weak augmentations, we use the same augmentations as those of the fully supervised and the weakly supervised setup. We evaluate different batch sizes (16, 32, 64 and 128) for this and train the models up to 300 epochs (for the best performing hyperparameter settings) using the ADAM optimizer and the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent Loss) [22]. We store the model checkpoints at every 50 epochs and compare their performances on the downstream abnormality detection task. After training the backbone encoder, we use it to obtain the representations for the normal and test data. The representations obtained from the normal class instances are then used to train an OC-SVM and the representations obtained from the new instances are then tested with the trained OC-SVM.

5.2 Evaluation metrics

This work focuses on the detection of abnormal instances (and not on the bag level decision). The extremely imbalanced task, where the number of abnormal instances is less than 2% of the total number of instances, requires special care with respect to the chosen performance metric. We envision the scenario of AI-supported abnormal instance detection, such as the detection of a few malignant cells among ten to fifty thousand cells on a WSI. We first conduct experiments and measure the performance of the models to distinguish abnormal from normal cells using *False Positive (FP)*, *False Negative (FN)*, *True Positive (TP)*, *Precision*, *Sensitivity* and *Specificity*. To most efficiently support the human expert (i.e., to present the cytologist with a small number of suspicious cells so that going through all the cells in the slide is not needed), detected abnormalities are typically presented in an ordered list with the most severe cases (in terms of abnormality of the cells, here based on the *confidence scores* of the models in inferring an instance as *abnormal*) first. A natural measure of performance is then to look at the ratio of true positives (TP) among the top K abnormal predicted instances. This leads us to the measure of *Recall@K*, commonly used in information retrieval.

$$\text{Recall@K} = \frac{\text{Number of Relevant Items in TopK}}{\text{Number of relevant items}} . \quad (3)$$

5.3 Training details

For the fully supervised and weakly supervised setup, we use a batch size of 64, learning rate of 0.001 and train the models for 30 epochs.

For the self-supervised training, we perform experiments using batch sizes of 16, 32, 64 and 128. We evaluate a number of temperature coefficients τ for the one-class training of the encoder and tolerance levels ν , kernel coefficients γ for the OC-SVM, and observed that $\tau = 2$, $\nu = 0.1$ and $\gamma = \text{"auto"}$ for the RBF kernel produce the best results.

6 Results

Here we summarize the performance of the models we trained on the held out test sets. Table 3 provides a condensed overview of the performance of the three evaluated approaches: fully supervised, weakly supervised, and OCC.

Supervised For the fully supervised setup, we observe in Table 3 high specificity and reasonable sensitivity on the test set. We can interpret the results of each binary classification as the model's ability to distinguish each abnormality when that abnormality class is present in the synthetic collection of cells which correspond to (in terms of number/size) a (moderate) WSI of a single sample. We observe that for the smaller abnormal classes (i.e. representing low witness

Table 3. Comparison of results for fully supervised, weakly supervised, and OCC setups. **FN**: False Negatives, **FP**: False Positives, **TP**: True Positives, **FS**: Fully Supervised, **WS**: Weakly Supervised

Abnormal Class	FP			FN			TP			Specificity			Sensitivity		
	FS	WS	OCC	FS	WS	OCC	FS	WS	OCC	FS	WS	OCC	FS	WS	OCC
BAS	17	3268	740	26	24	7	107	109	8	99.78%	58.49%	90.6%	80.45%	81.95%	53.33%
HAC	2	2819	740	66	59	2	57	64	1	99.97%	64.19%	90.6%	46.34%	52.03%	33.33%
OTH	11	4422	740	11	2	52	78	87	37	99.86%	43.83%	90.6%	87.64%	97.75%	41.57%
LYI	0	4040	740	14	7	106	6	13	17	100%	100%	90.6%	30%	30%	13.82%
FGC	0	2978	740	3	4	9	12	11	4	100%	62.17%	90.6%	80%	73.33%	30.77%
KSC	0	4480	740	0	4	110	13	9	33	100%	43.10%	90.6%	100%	69.23%	23.08%
ABE	0	4213	740	3	3	14	0	0	6	100%	46.49%	90.6%	0%	0%	30%

Sensitivity for Different Batch Sizes and Outlier Classes for the OCC Setup

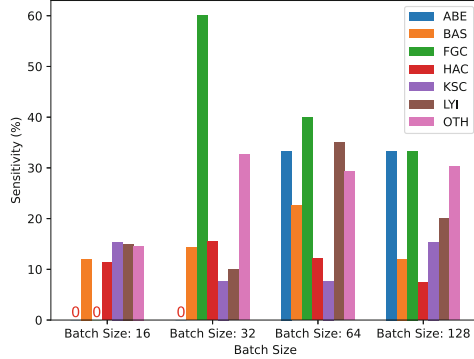


Fig. 2. Sensitivity for Different Batch Sizes and Outlier Classes

rate in a whole slide cytology image), the sensitivity sometimes drops dramatically. The average specificity and sensitivity for all the seven outlier classes are **99.95%** and **60.63%**.

Weakly supervised As can be seen in Table 3, the number of false positives in the weakly supervised case is very high for all the seven classes. The average specificity and sensitivity for all the seven outlier classes are **52.42%** and **62.76%**.

One-class classification We observed superior downstream abnormal cell detection performance for the `CenterCropResize` augmented training dataset and thus report here the performances (related to this augmentation) for different batch sizes when trained for 150 epochs and also for model checkpoints saved at every 50 epochs up to 300 for the best performing batch size. We also report the framework’s best performance when trained for 100 epochs with batch size 64 in Table 3.

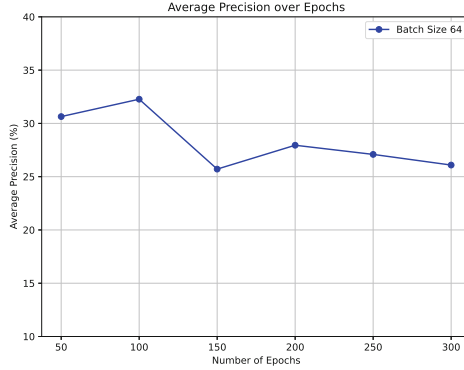


Fig. 3. Average Precision over epochs for batch size 64

We present the summary of the One-Class Classification (OCC) sensitivities as a bar plot for different batch sizes (when trained for 150 epochs) in Figure 2. We also report their average (across all the 7 outlier classes) precision in Table 4. We observe a roughly constant specificity of 90% for all the models.

Table 4. Average Precision for Different Batch Sizes when trained for 150 epochs

Batch Size	16	32	64	128
Average Precision (%)	9.77	20.00	25.71	21.68

We further evaluate model checkpoints saved at every 50 epochs for the best performing batch size of 64. The variation of average precision for the model with batch size 64 with the number of epochs is shown in Figure 3. We observe that the best performing model is the one saved at epoch 100.

6.1 Recall@K

Although in Table 3 we observe the OCC method faring poorly on the whole dataset level, we envision this method to be used as a ranking-based recommendation system for the detection of abnormal cells by cytologists. To evaluate its performance in that task, we compute the $Recall@K$ for all three setups: *fully supervised*, *weakly supervised* and *one class classification* for all seven outlier classes. We compute $Recall@K$ for $K = 400$ (which can be presented as a 20×20 grid of abnormal predicted cells to the cytologist) and observe that the performance for the *One-Class Classification* setup performs better than the *Weakly Supervised* setup. We present the $Recall@400$ scores for the 3 different setups as a bar plot for all the 7 outlier classes in Figure 4. We also show the *number of TPs* in Table 6, where we can observe that except for the *BAS*

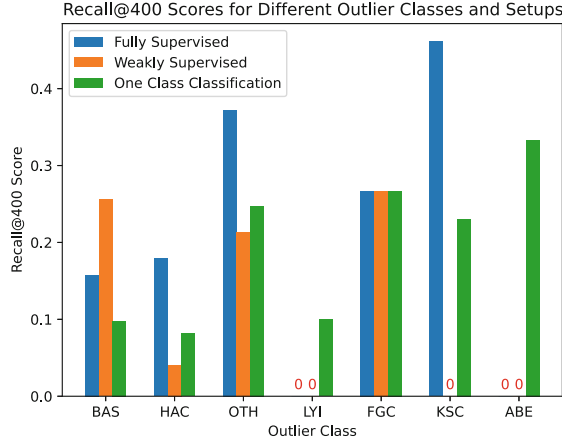


Fig. 4. Recall@400 for all the three setups for the 7 outlier classes

class, for all the other outlier classes, the OCC method performed better than the weakly supervised one in identifying the abnormality. Table 5 shows that the average *Recall@400* score for OCC surpasses that of the Weakly Supervised setup, however (not surprisingly) lags slightly behind that of the Fully Supervised setup (which has access to instance-level annotations – unavailable in the real scenario).

Table 5. Average Recall@400 across all Outlier Classes for different setups

Fully Supervised	Weakly Supervised	OCC
0.2051	0.1109	0.1938

Table 6. The number of TPs in the top 400 for different setups and Outlier Classes

	BAS	HAC	OTH	LYI	FGC	KSC	ABE
Fully Sup.	21	22	33	00	04	06	00
Weak Sup.	34	05	19	00	04	00	00
OCC	13	10	22	02	04	03	01

7 Discussion and Conclusion

Our results indicate that one-class classification may be a good way to handle the case of detecting key instances (even when there are very is few of them) in very large bags; a situation not uncommon in healthcare diagnostics. This can facilitate the detection of early stage cancers where the number of cancer cells

present in a cytology slide might be extremely low and might be missed by the cytologist during inspection.

In particular, the **CenterCropResize** augmentation effectively reduces the number of false positives (now roughly 10% of the number of samples as compared to as much as 60% for the weakly-supervised one), whose value is also dependent on the tolerance level chosen for the OC-SVM. Further, it can be inferred that, if choosing a proper batch-size and tolerance value, the OC-SVM based on contrastively learned representations has the potential to surpass the weakly-supervised learning framework.

Acknowledgements. This work is supported by Sweden’s Innovation Agency (VINNOVA) projects 2017-02447, 2020-03611, 2021-01420, Swedish Cancer Society (Cancerfonden) projects 22 2353 Pj and 222357 Pj., and the Centre for Interdisciplinary Mathematics (CIM), Uppsala University.


References

1. Aeffner, F., Zarella, M.D., Buchbinder, N., Bui, M.M., Goodman, M.R., Hartman, D.J., Lujan, G.M., Molani, M.A., Parwani, A.V., Lillard, K., et al.: Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *Journal of pathology informatics* **10**(1), 9 (2019)
2. Andersson, A., Koriakina, N., Sladoje, N., Lindblad, J.: End-to-end multiple instance learning with gradient accumulation. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 2742–2746. IEEE (2022)
3. Chong, P., Ruff, L., Kloft, M., Binder, A.: Simple and effective prevention of mode collapse in deep one-class classification. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2020)
4. Eccher, A., Girolami, I.: Current state of whole slide imaging use in cytopathology: pros and pitfalls. *Cytopathology* **31**(5), 372–378 (2020)
5. Edman, K., Stark, C.R., Basic, V., Lindblad, J., Hirsch, J.M.: Dental hygienists and dentists as providers of brush biopsies for oral mucosa screening. *Int. J. Dental Hygiene* **21**(3), 524–532 (2023)
6. George, A., Sreenivasan, B., Sunil, S., Varghese, S.S., Thomas, J., Gopakumar, D., Mani, V.: Potentially malignant disorders of oral cavity. *Oral Maxillofac Pathol J* **2**(1), 95–100 (2011)
7. 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)
8. Hirsch, J.M., Sandy, R., Hasséus, B., Lindblad, J.: A paradigm shift in the prevention and diagnosis of oral squamous cell carcinoma. *Journal of Oral Pathology & Medicine* **52**(9), 826–833 (2023)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
10. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. *Adv. Neural. Inf. Process. Syst.* **33**, 20578–20589 (2020)
11. Koriakina, N., Sladoje, N., Bašić, V., Lindblad, J.: Deep multiple instance learning versus conventional deep single instance learning for interpretable oral cancer detection. *PLoS ONE* **19**(4), e0302169 (2024)

12. Koriakina, N., Sladoje, N., Lindblad, J.: The effect of within-bag sampling on end-to-end multiple instance learning. In: 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA). pp. 183–188. IEEE (2021)
13. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
14. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**(6), 555–570 (2021)
15. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. *Advances in neural information processing systems* **10** (1997)
16. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T., Marr, C.: An expert-annotated dataset of bone marrow cytology in hematologic malignancies [data set]. *Cancer Imaging Arch* (2021)
17. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems* **32** (2019)
18. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International conference on machine learning. pp. 4393–4402. PMLR (2018)
19. Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., Khandeparkar, H.: A theoretical analysis of contrastive unsupervised representation learning. In: International Conference on Machine Learning. pp. 5628–5637. PMLR (2019)
20. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. *Advances in neural information processing systems* **12** (1999)
21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural. Inf. Process. Syst.* **34**, 2136–2147 (2021)
22. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems* **29** (2016)
23. Sohn, K., Li, C.L., Yoon, J., Jin, M., Pfister, T.: Learning and evaluating representations for deep one-class classification. [arXiv:2011.02578](https://arxiv.org/abs/2011.02578) (2020)
24. Tsybakov, A.B.: Nonparametric estimators. In: Springer Series in Statistics, pp. 1–76. Springer series in statistics, Springer New York, New York, NY (2009)
25. Walsh, T., Liu, J.L., Brocklehurst, P., Glenney, A.M., Lingen, M., Kerr, A.R., Ogden, G., Warnakulasuriya, S., Scully, C.: Clinical assessment to screen for the detection of oral cavity cancer and potentially malignant disorders in apparently healthy adults. *Cochrane Database of Systematic Reviews* (11) (2013)
26. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
27. Zhai, S., Cheng, Y., Lu, W., Zhang, Z.: Deep structured energy based models for anomaly detection. In: International conference on machine learning. pp. 1100–1109. PMLR (2016)
28. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations (2018)



Hybrid CNN-LSTM Framework for Enhanced Congestive Heart Failure Diagnosis: Integrating QRS Detection

Aditya Oza^(✉) , Sanskriti Patel , Bhavesh Gyanchandani , Abhinav Roy ,
and Santosh Kumar 

IIIT, Naya, Raipur, India
aditya21102@iiitnr.edu.in

Abstract. Our research tackles the critical issue of congestive heart failure (CHF), a serious cardiovascular condition in which the heart's ability to pump blood effectively is compromised, leading to fluid buildup. Existing diagnostic methods often struggle with signal processing and manual Electrocardiogram (ECG) analysis, resulting in reduced accuracy and added complexity in diagnosis. To overcome these challenges, we present an innovative framework that integrates QRS detection of RR peaks and intervals from ECG data. We then propose a hybrid CNN-LSTM network specifically designed for CHF diagnosis. What sets our approach apart is the strategic application of deep learning, combining the feature extraction strengths of Convolutional Neural Networks (CNNs) with the temporal processing capabilities of LSTM networks. This combination enhances diagnostic outcomes, significantly improving the accuracy of CHF detection and representing a notable advancement in clinical diagnostics. Our research highlights the potential of deep learning to enhance diagnostic precision and support clinical decision-making for CHF. By leveraging advanced technologies and methodologies, we aim to revolutionize cardiovascular health monitoring and contribute to more effective patient care strategies. This innovative approach achieves an impressive accuracy rate of 98.77

Keywords: Congestive Heart Failure · Electrocardiogram · Convolutional Neural Networks · Long Short-Term Memory

1 Introduction

Congestive heart failure (CHF) is a prevalent cardiovascular condition characterized by the heart's diminished ability to effectively pump blood, resulting in fluid accumulation and various complications [1]. In a healthy heart, efficient stroke volume-blood flow volume ejected per heartbeat-ensures oxygen-rich blood is adequately distributed throughout the body from the left ventricle. However, in CHF, particularly with impaired pump function, stroke volume decreases. The heart undergoes remodeling, becoming enlarged with stiffened muscle walls

stretched to accommodate more oxygen-rich blood, leading to decreased pumping efficiency and increased susceptibility to fatigue [1]. This condition also causes blood and fluid to accumulate in the lungs and body, resulting in symptoms like breathlessness and generalized swelling [1].

Diagnosing CHF is primarily clinical, relying on a comprehensive assessment of symptoms, signs, and corroborative evidence from diagnostic tests [1]. Among these tests, the electrocardiogram (ECG) plays a pivotal role as a noninvasive tool to record and analyze the heart's electrical activities [1]. While ECG signals are known to exhibit alterations in CHF patients, these changes are often non-specific and insufficient for definitive diagnosis through standard manual analytic methods alone. Typically, cardiologists visually inspect ECG readings to identify abnormalities, a process prone to time-consuming evaluations and inter-observer variability [1].

Our research addresses these diagnostic challenges by proposing an innovative framework that integrates advanced signal processing techniques with deep learning (DL) methodologies. Specifically, we focus on enhancing CHF detection accuracy through a hybrid CNN-LSTM network tailored for analyzing ECG data. This approach combines the robust feature extraction capabilities of CNN with the sequential learning strengths of Long Short-Term Memory (LSTM) networks, optimizing the identification of subtle patterns indicative of CHF.

By leveraging DL techniques, our framework aims to revolutionize cardiovascular health monitoring, improving diagnostic precision and facilitating early intervention strategies. Our study contributes to the evolving landscape of medical diagnostics by demonstrating significant advancements in CHF detection accuracy and clinical decision-making. We achieve a high accuracy rate of 98.77% in our evaluations, underscoring the practical implications of integrating CNN-LSTM fusion techniques for enhancing diagnostic outcomes in cardiovascular care.

1.1 Contributions

The major contributions of our research are as follows:

1. Proposal of an innovative framework integrating CNN-LSTM networks for enhanced CHF detection accuracy using ECG data.
2. Application of advanced signal processing techniques to improve feature extraction and anomaly detection in ECG signals related to CHF.
3. Validation of our framework through extensive empirical evaluations, demonstrating superior diagnostic accuracy compared to existing methods.
4. Contribution to the field of medical diagnostics by leveraging DL methodologies to advance cardiovascular health monitoring and early intervention strategies.

In this paper, we present a detailed methodology and empirical results to validate the effectiveness of our proposed framework. We begin with a comprehensive review of related work in Section II, highlighting current challenges

and gaps in existing diagnostic approaches. Section III outlines our methodology, encompassing data preprocessing techniques, the architecture of the CNN-LSTM model, and evaluation metrics. Results and discussions are presented in Section IV, followed by conclusions and avenues for future research in Section V.

2 Literature Survey

Savarese G, et al. [1] offer a thorough review of key advancements in heart condition diagnosis, with a particular emphasis on CHF. L. Zou et al. [2] introduced a novel architecture combining LSTM networks with Deep Convolutional Neural Networks (DCNN), achieving a real-time CHF detection accuracy of 97.62

Shrivastava et al. [4] propose a robust approach for detecting myocardial infarction (MI) by utilizing three feature selection methods and eight machine learning (ML) algorithms. Their evaluation with standardized data demonstrates superior predictive capabilities, outperforming existing studies in the field. Similarly, studies [5–7] incorporate federated learning with an RF approach, achieving a 95% diagnostic accuracy for CHF. The HBA-FRCNN technique specifically addresses ECG signal noise artifacts, reaching a 97.65% accuracy in chronic heart failure prediction.

Rai et al. [8] examined various ML and DL methods for cardiac disease detection, providing insights into the application of cutting-edge techniques for medical diagnosis. Khan et al. [9] conducted a comprehensive analysis of current ML models for predicting cardiac arrests, emphasizing the need for rigorous evaluation to improve healthcare prediction accuracy. Their findings highlight the importance of enhanced threat evaluation techniques to improve outcomes and optimize resource allocation in cardiovascular disease (CVD) diagnosis.

Bhaskarpandit et al. [10] showcased significant advancements in cardiac diagnosis through their research on eigendomain deep representation learning for analyzing 12-lead ECG trace images for MI diagnosis. Their work highlights the potential of ML and DL techniques to enhance diagnostic accuracy and patient care. This study provides a foundational understanding of ECG signal acquisition and heart disease detection.

The literature review identifies several gaps in the existing research:

1. Limited comparative studies on ML algorithms for early CHF detection using ECG data, and insufficient exploration of fusion techniques for improved predictive accuracy.
2. Inadequate research on the scalability and adaptability of DL techniques for CHF detection across diverse patient groups and healthcare settings.
3. Minimal investigation into the effects of data preprocessing methods on DL model performance in CHF prediction, especially in reducing ECG signal noise and artifacts.
4. Lack of focus on potential biases affecting DL model generalization, and insufficient studies on model interpretability to enhance clinical decision-making and adoption.

Savarese et al. [1] reviewed significant advancements in CHF diagnosis, emphasizing deep learning (DL) methodologies, while Zou et al. [2] introduced an architecture integrating LSTM and DCNN for real-time CHF detection, demonstrating improved accuracy. Our study builds on this by leveraging DL's strengths in hierarchical feature extraction and temporal modeling to enhance CHF detection from ECG signals, focusing on robust preprocessing to mitigate noise and artifacts. This approach advances clinical diagnostics, highlights DL's crucial role in CHF detection, and underscores the need for further research to address existing gaps and enhance healthcare applications.

3 Methods and Materials

To achieve early CHF detection based on ECG data [10] [11], we propose a comprehensive methodology, starting with dataset description and data preprocessing to standardize recordings, reduce noise, and enhance signal quality. Key patterns indicative of CHF are captured using feature extraction techniques. Our hybrid CNN-LSTM model, central to this approach, combines the feature extraction power of CNNs with LSTMs' temporal processing strengths to distinguish between healthy and CHF-affected ECG signals. Figure 5 illustrates the workflow from preprocessing to CHF classification. The model's impact on clinical decision-making is evaluated through performance metrics, highlighting its potential to improve healthcare diagnostics with advanced machine learning techniques.

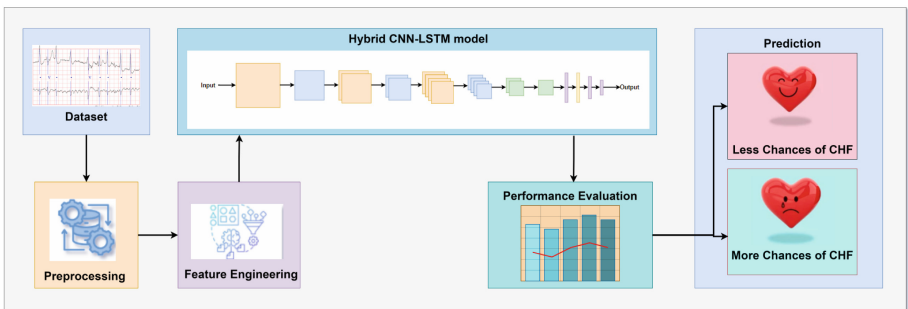


Fig. 1. Overview of the approach for early detection of CHF based on ECG data. The process includes dataset preprocessing, feature extraction, and classification using a hybrid CNN-LSTM model.

Referring to Algorithm 18, the methodology outlines the step-by-step process for early detection of CHF based on ECG data using the hybrid CNN-LSTM model.

Algorithm 1 Early Detection of CHF Based on ECG Data

```

1: Input: ECG dataset  $D$ 
2: Output: CHF predictions
3: procedure CHF DETECTION( $D$ )
4:   Initialization:
5:   Load dataset  $D = \{X, Y\}$  and perform preprocessing
6:   Split  $D$  into training and testing sets
7:   Data Preprocessing:
8:   Standardize  $X$ :  $X_{std} = \frac{X - \mu}{\sigma}$ 
9:   Reduce noise:  $X_{noise\_reduced} = \text{NoiseReduction}(X_{std})$ 
10:  Normalize:  $X_{norm} = \frac{X_{noise\_reduced} - \min(X_{noise\_reduced})}{\max(X_{noise\_reduced}) - \min(X_{noise\_reduced})}$ 
11:  Feature Extraction:
12:  Extract key features:  $QRS_{Duration}, IBI$ 
13:  Model Training:
14:  Train hybrid CNN-LSTM:  $M_{CNN-LSTM} = \text{TrainHybridCNNLSTM}(X_{norm}, Y)$ 
15:  Predictions and Evaluation:
16:  Predict CHF:  $\hat{Y} = \text{Predict}(M_{CNN-LSTM}, X_{test})$ 
17:  Evaluate model performance
18: end procedure

```

3.1 Dataset Description

In this research, we use the BIDMC CHF and MIT-BIH datasets [12, 13] dataset, ensuring the quality of our training dataset is vital for our model’s effectiveness. We describe a systematic approach for data collection and preprocessing, aligning with established best practices. Our methods draws data from the BIDMC and MIT-BIH datasets, with preprocessing techniques applied to transform raw ECG recordings. By leveraging insights from leading cardiovascular health research, we aim to enhance the generalizability and dependability of our model.

Table 1. Databases used in our research

Database	NHYA Class	# Subjects	# Males (age)	# Females (age)	# EB
BIDMC	CHF	15	11(22 – 71 years)	4(54 – 63 years)	20,000
MIT-BIH	NSR	18	5(26 – 45 years)	13(20 – 50 years)	36,000

Abbreviation: EB=Extracted beats

BIDMC CHF Database [11] BIDMC CHF Database [11]: Contains ECG recordings from 15 subjects (11 males aged 22-71, 4 females aged 54-63) with severe CHF (NYHA class 3-4). These recordings provide valuable insights into cardiac dynamics in severe heart failure cases.

MIT-BIH NSRDB [12] Comprises 18 recordings from subjects without significant arrhythmias, including 13 females (20-50 years) and 5 males (26-45 years).

Each 20-hour recording contains two ECG signals sampled at 250 Hz with 12-bit resolution over ± 10 mV. Annotations were generated by an automatic detector.

Total Heartbeats The combined datasets yield a total of 490,505 heartbeats, forming the basis for our predictive model development and evaluation.

3.2 Data Preprocessing

Preprocessing ECG data is crucial for ensuring the accuracy and reliability of subsequent analysis. This study employs a rigorous standardization process to address discrepancies, missing data, and noise in ECG recordings, ensuring consistency and enhancing the reliability of the analysis [1, 14, 15]. The first step involves noise reduction, specifically targeting variations in non-QRS signals. This is achieved using two Moving Average Cascades (MACs) with different impulse response intervals—140 ms for the wider MAC and 25 ms for the narrower one—aimed at preserving QRS complex peaks while attenuating slow waves like T waves and baseline drifts [2].

The second step enhances the QRS complexes by applying derivative filters. During initialization, the filter that maximizes a signal quality index (SQI) is selected based on the following equation:

$$\text{SQI} = \frac{k_s + mD_s}{k_n + mD_n}$$

Here, mD_s represents the trimmed mean of maxima within 1.6-second windows, capturing high derivative values specific to QRS complexes, while mD_n represents the trimmed mean within 0.09-second windows to account for noise. The filter with the highest SQI, denoted as MAC2[n], is applied to all ECG records [6].

The third phase involves decision reasoning for QRS event detection. The output of the derivative filter is compared to an adaptive threshold, updated dynamically to maintain accuracy while minimizing false positives:

$$T = \begin{cases} \min(\bar{D}, 2.5 \cdot T^{(0)}) \\ \max(\bar{D}, 0.5 \cdot T^{(0)}) \end{cases}$$

The threshold is adjusted based on the distance from the previous QRS detection, ensuring flexibility in response to changing signal conditions. The position of the maximum (or minimum) of the derivative signal is used to accurately place the fiducial point for each detected QRS complex.

Finally, the extracted features undergo normalization using MinMaxScaler to standardize their scales, which is crucial for avoiding biases during model training. Scatter plots visualize the normalized data, offering insights into the distribution of characteristics distinguishing CHF from normal conditions [4]. Figure 2 illustrates how normalization helps standardize the data.

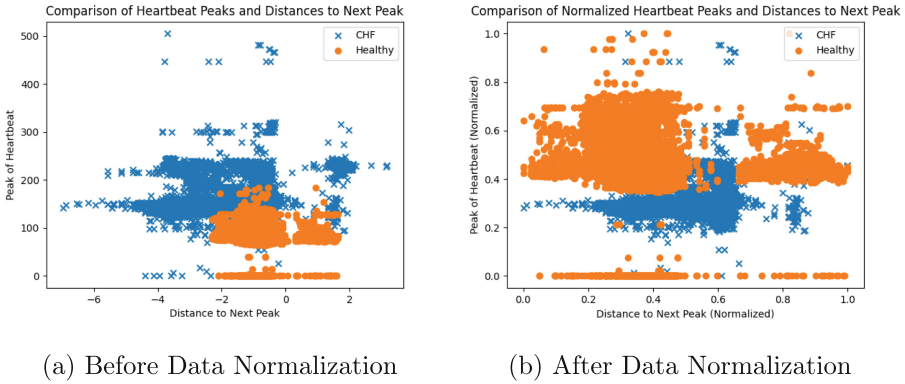


Fig. 2. Diagram represents the heartbeat comparison (a) Before Data Normalization and (b) After Data Normalization.

3.3 Feature extraction

Feature engineering is a crucial step in extracting meaningful information from the ECG signals [16–18]. The following key features are engineered to capture essential aspects of cardiac dynamics:

QRS Wave The GQRS detection technique plays a crucial role in our investigation, accurately identifying QRS complexes in ECG waveforms. This precision is essential for meaningful feature extraction and reliable predictive models for early CHF detection. Rigorous validation confirms the technique’s suitability and effectiveness. Figure 3 shows the ECG after the GQRS detection.



Fig. 3. Corrected GQRS R-Peak Detection used to extract R Peaks and RR Intervals

$$Sp = \frac{\text{CorrectQRSPredicted}}{\text{TotalnumberoftrueQRSpeaks}}$$

$$Pp = \frac{\text{CorrectQRSPredicted}}{\text{TotalnumberofQRSPredicted}}$$

$$F1 = 2 \times \frac{Sp \times Pp}{Sp + Pp}$$

The QRS complex represents the depolarization of the ventricles and is a crucial feature in ECG analysis. Its duration (*QRS_Duration*) can be calculated as the time taken from the onset to the offset of the QRS complex:

$$QRS_Duration = QRS_Offset - QRS_Onset$$

The Inter-Beat Interval, also known as the RR interval, reflects the period between two R-peaks. It is a fundamental measure of heart rate variability (*HRV*) and is computed as:

$$IBI = R_n - R_{n-1}$$

where R_n and R_{n-1} are the locations of consecutive R-peaks.

To ensure consistency in feature scaling, normalization is applied to the extracted features. The Min-Max normalization is employed:

$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

where X represents the value of the feature, and $\min(X)$ and $\max(X)$ are the min and max values of the feature, respectively. Figure 4: This figure (a) and (b) provide a comparison of the distribution of RR-Intervals and R Peaks, respectively. Understanding these distributions is crucial as they represent key features in ECG data that can help in analyzing heart rhythm patterns and detecting abnormalities.

4 Proposed Hybrid CNN-LSTM Network Architecture

We propose a hybrid DL model combining CNN and LSTM networks for early CHF detection. The architecture of our hybrid CNN-LSTM model is detailed in Figure 5

The hybrid CNN-LSTM model architecture is designed to leverage the spatial hierarchical features learned by the CNN layers and the temporal dependencies

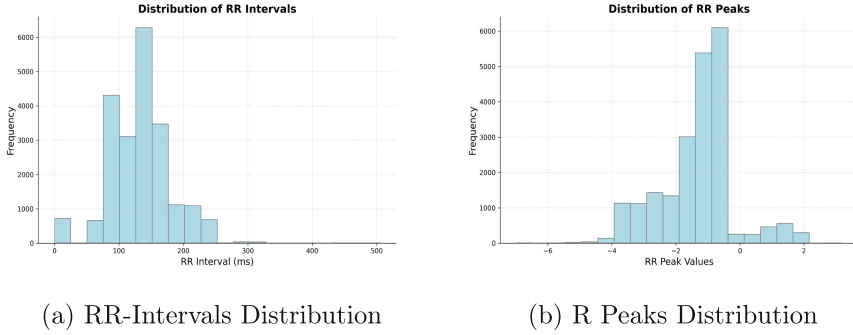


Fig. 4. Diagram represents the distribution comparison (a) RR-Intervals distribution and (b) R Peaks distribution.

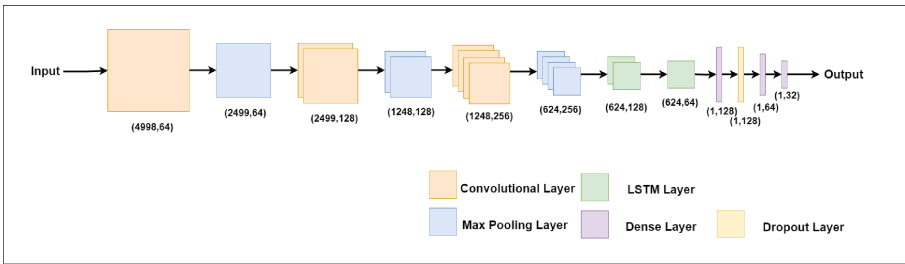


Fig. 5. Proposed Hybrid CNN-LSTM Network Architecture

captured by the LSTM layers. The CNN layers perform feature extraction from the input ECG signals, capturing both local and global patterns. These features are then fed into the LSTM layers, which utilize their sequential learning ability to model long-term dependencies in the ECG data.

4.1 Mathematical Formulation

The operations performed by the CNN layers can be represented as:

$$y_{\text{CNN}} = \text{CNN}(\mathbf{X}; \theta_{\text{CNN}})$$

where \mathbf{X} denotes the input ECG signals, y_{CNN} represents the output features extracted by the CNN layers parameterized by θ_{CNN} .

The LSTM layers subsequently process these extracted features:

$$y_{\text{LSTM}} = \text{LSTM}(y_{\text{CNN}}; \theta_{\text{LSTM}})$$

where y_{LSTM} represents the output of the LSTM layers parameterized by θ_{LSTM} .

The final output of the hybrid model is obtained by applying a classification layer on top of the LSTM outputs:

$$\hat{y} = \text{softmax}(y_{\text{LSTM}}; \theta_{\text{class}})$$

where $\hat{\mathbf{y}}$ represents the predicted probabilities of different classes (e.g., normal vs. CHF) and Θ_{class} denotes the parameters of the classification layer.

4.2 Strengths and Uniqueness

Our proposed hybrid CNN-LSTM model offers several strengths and unique advantages:

1. **Spatial-Temporal Integration:** Combines CNN for spatial features and LSTM for temporal modeling, capturing both local and global ECG patterns.
2. **Improved Accuracy:** Enhances early CHF detection accuracy by leveraging CNN and LSTM strengths.
3. **Longitudinal Data Handling:** LSTM layers effectively process time-series data, suitable for monitoring CHF progression.
4. **Scalability and Efficiency:** Maintains computational efficiency for real-time clinical applications despite model complexity.

The proposed architecture aims to effectively capture both spatial and temporal characteristics of ECG signals, thereby enhancing the accuracy of CHF detection.

4.3 Model Architecture Summary

4.4 Impact and Applications

The proposed predictive model for early detection of CHF based on ECG data holds significant implications for clinical practice and public health.

Clinical Impact The accurate identification of CHF at an early stage enables proactive clinical interventions, resulting in better patient outcomes and lower medical expenses. The model's precision in distinguishing between CHF and NSR cases contributes to timely and targeted medical interventions.

Quantifying Impact The impact of our model can be quantified using metrics such as:

$$\text{CS (\%)} = \frac{\text{Cost without model} - \text{Cost with model}}{\text{Cost without model}} \times (100)$$

cost Saving (CS) where the cost includes expenses related to late-stage CHF treatments, hospitalizations, and emergency care.

5 Results and Discussion

5.1 Accuracy and Loss

The training and validation accuracy and loss curves provide insights into the model's learning process and its ability to generalize to unseen data.

The accuracy and loss curves indicate that the model converges well during training, with minimal overfitting. The hybrid CNN-LSTM architecture ensures that the model captures both spatial and temporal features, contributing to its high accuracy and low loss values.

Table 2. Detailed Summary of the Hybrid CNN-LSTM Model Architecture

Layer Type	Output Shape	Param #	Description
Input Layer	(None, 5000, 1)	-	Input ECG signals
Convolutional Layer 1	(None, 4998, 1, 64)	256	Filters: 64, Kernel: 3x3
Max Pooling Layer 1	(None, 2499, 1, 64)	-	Pool Size: 2x2
Convolutional Layer 2	(None, 2497, 1, 128)	24,704	Filters: 128, Kernel: 3x3
Max Pooling Layer 2	(None, 1248, 1, 128)	-	Pool Size: 2x2
Convolutional Layer 3	(None, 1246, 1, 256)	98,560	Filters: 256, Kernel: 3x3
Max Pooling Layer 3	(None, 623, 1, 256)	-	Pool Size: 2x2
LSTM Layer 1	(None, 623, 128)	197,120	Hidden Units: 128
LSTM Layer 2	(None, 623, 64)	49,408	Hidden Units: 64
Dense Layer 1	(None, 128)	8,320	Dense: 128
Dropout Layer	(None, 128)	-	Dropout
Dense Layer 2	(None, 64)	8,256	Dense: 64
Dense Layer 3	(None, 32)	2,080	Dense: 32
Output Layer	(None, Number of classes)	-	Output: Number of classes

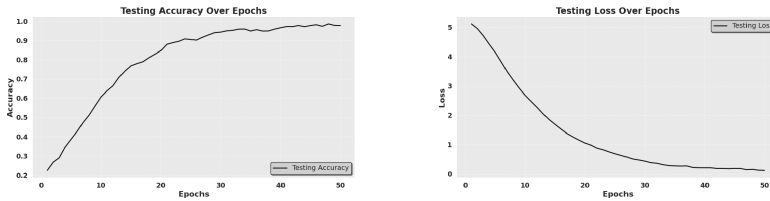


Fig. 6. Training and Validation Accuracy and Loss Curves

5.2 Confusion Matrix and Precision-Recall Curve

The confusion matrix (Figure 7) provides a detailed breakdown of the model’s classification performance, crucial for understanding its effectiveness in distinguishing between CHF and Normal Sinus Rhythm (NSR) segments.

The confusion matrix shows a significant number of True Positives (TP = 2159) and True Negatives (TN = 4090), with only a few False Positives (FP = 47) and False Negatives (FN = 31). These results highlight the model’s strong ability to accurately classify both CHF and NSR segments, demonstrating high precision and recall.

The Precision-Recall (PR) curve (Figure 8) and its AUC of 0.989 further highlight the model’s capability to maintain high precision and recall across different thresholds. The hybrid CNN-LSTM model’s ability to extract and utilize both spatial and temporal features is a key factor in achieving such high performance metrics.

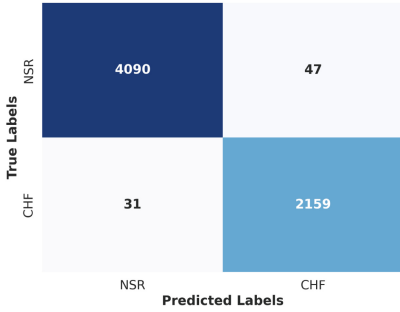


Fig. 7. Confusion Matrix of the Hybrid CNN-LSTM Model

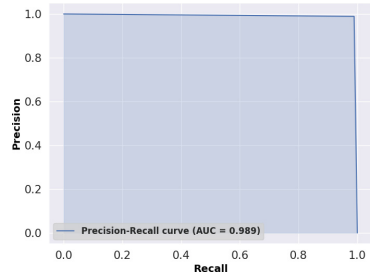


Fig. 8. Precision-Recall Curve of the Hybrid CNN-LSTM Model

5.3 Classification Report

The classification report (Table 3) provides detailed metrics such as precision, recall, and F1-score for both classes (NSR and CHF).

Table 3. Classification Report for the Hybrid CNN-LSTM Model

Class	Precision	Recall	F1-Score	Support
NSR	0.978	0.988	0.983	4137
CHF	0.979	0.978	0.979	2206
Avg/Total	0.978	0.978	0.978	6343

The high precision, recall, and F1-scores across both classes demonstrate the robustness of the hybrid CNN-LSTM model. The CNN layers effectively extract meaningful features from the ECG data, while the LSTM layers model the sequential nature of the data, leading to superior classification performance.

5.4 Discussion

The results demonstrate the effectiveness of the proposed hybrid CNN-LSTM model in accurately detecting CHF from ECG data. The model’s architecture, which combines CNN’s spatial feature extraction and LSTM’s temporal sequence learning, is pivotal in achieving high classification performance. This hybrid approach not only improves accuracy but also ensures robust detection of CHF, making it a valuable tool for clinical diagnostics and early intervention.

The combination of spatial and temporal feature extraction allows the model to capture both short-term patterns (through CNN) and long-term dependencies (through LSTM), thereby enhancing its ability to discriminate between NSR and

CHF segments. This capability is crucial in medical applications where both immediate and prolonged ECG characteristics play a role in diagnosis.

Furthermore, the high AUC-PR indicates that the model maintains high precision and recall even when the threshold for classifying CHF segments varies. This flexibility is essential in clinical settings where different decision thresholds may be required based on the specific diagnostic needs.

6 Comparative Analysis

We compare our proposed Hybrid CNN-LSTM method with several existing approaches as summarized in Table 4. Each study utilized different datasets and methodologies for cardiac arrhythmia detection, achieving varying levels of accuracy. Our method, using MIT-BIH and BIDMC CHF datasets, achieves an accuracy of 98.77%.

The studies listed in Table 4 employ a variety of techniques, including Unet++, Faster RCNN classifiers, DA-DRRNet, Eigendomain DRL approaches, and Artificial Neural Networks. These methods have shown commendable accuracies ranging from 89.83% to 98.68%.

Our choice of Hybrid CNN-LSTM is particularly effective due to its ability to harness both convolutional and LSTM layers. This architecture enables the model to capture intricate temporal patterns present in ECG signals. By integrating spatial and sequential learning, our approach excels in detecting complex cardiac arrhythmias, as evidenced by its high accuracy across different datasets.

Table 4. Result Analysis

Ref.	Dataset used	Methodology used	Accuracy
[2]	NSR-RR, CHF-RR	Unet++	89.83%
[6]	BIDMC-CHF, MITBIH	Faster RCNN classifier	98%
[8]	BIDMC-CHF, PTBDB	DA-DRRNet	98.57%
[9]	Mendeley data source	Eigendomain DRL approach	98.68%
[12]	from Catholic University of Leuven	Artificial Neural Network	90.00%
[18]	own dataset	ML algorithms	94.00%
Proposed	MIT-BIH, BIDMC CHF	Hybrid CNN-LSTM	98.77%

7 Conclusion

In this study, we introduced a Hybrid CNN-LSTM model for the detection of cardiac arrhythmias using ECG signals from MIT-BIH and BIDMC CHF datasets. Our model achieved a high accuracy of 98.77%, showcasing its effectiveness in accurately classifying normal sinus rhythm (NSR) and CHF segments.

The Hybrid CNN-LSTM architecture proved advantageous due to its ability to capture both spatial features through convolutional layers and temporal dependencies through LSTM layers. This dual capability is crucial for handling the complex temporal patterns inherent in ECG signals.

Through a comparative analysis with existing methodologies, we demonstrated that our approach offers a robust solution to cardiac arrhythmia detection. The combination of convolutional and LSTM layers allows our model to excel in capturing intricate patterns, leading to superior performance in comparison to other methods.

References

1. Savarese G et al. Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovasc Res.* 2023 Jan 18;118(17):3272-3287. <https://doi.org/10.1093/cvr/cvac013>. Erratum in: *Cardiovasc Res.* 2023 Jun 13;119(6):1453. PMID: 35150240
2. L. Zou, et al. "Automatic Detection of Congestive Heart Failure Based on Multiscale Residual UNet++: From Centralized Learning to Federated Learning," in *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-13, 2023, Art no. 4001013, <https://doi.org/10.1109/TIM.2022.3227955>
3. Malik A, et al. Congestive Heart Failure. [Updated 2023 Nov 5]. In: *StatPearls* [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK430873/>
4. A. Shrivastava et al., "A Novel Hybrid Model for Predictive Analysis of Myocardial Infarction using Advanced Machine Learning Techniques," 2023 OITS International Conference on Information Technology (OCIT), Raipur, India, 2023, pp. 381-386, <https://doi.org/10.1109/OCIT59427.2023.10430780>
5. Ning, et al.: Automatic detection of congestive heart failure based on a hybrid deep learning algorithm in the internet of medical things. *IEEE Internet Things J.* **8**(16), 12550–12558 (2020)
6. S. Irin Sherly et al., An efficient honey badger based Faster region CNN for chronic heart Failure prediction, *Biomedical Signal Processing and Control*, Volume 79, Part 2, 2023
7. Baral, et. al. "A Literature Review for Detection and Projection of Cardiovascular Disease Using Machine Learning." *EAI Endorsed Transactions on Internet of Things* 10 (2024)
8. Prabhakararao, E., et al.: Congestive Heart Failure Detection From ECG Signals Using Deep Residual Neural Network. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **53**(5), 3008–3018 (2023). <https://doi.org/10.1109/TSMC.2022.3221843>
9. Bhaskarpandit, et al.: Detection of Myocardial Infarction From 12-Lead ECG Trace Images Using Eigendomain Deep Representation Learning. *IEEE Trans. Instrum. Meas.* **72**, 1–12 (2023)
10. Rani et al. (2024). An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction. *Archives of Computational Methods in Engineering*, 1-19
11. A. S et al. "ECG Classification and Arrhythmia Detection Using Wavelet Transform and Convolutional Neural Network," 2021 International Conference on Communication, Control and Information Sciences (ICCISc), Idukki, India, 2021, pp. 1-5, <https://doi.org/10.1109/ICCISc52257.2021.9485012>

12. D. Bibicu et al, "Cardiac Cycle Phase Estimation in 2-D Echocardiographic Images Using an Artificial Neural Network," in *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1273-1279, May 2013, <https://doi.org/10.1109/TBME.2012.2231864>
13. Goldberger, et al.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*. **101**(23), e215–e220 (2000)
14. Ju, R., et al.: 3D-CNN-SPP: A Patient Risk Prediction System From Electronic Health Records via 3D CNN and Spatial Pyramid Pooling. *IEEE Transactions on Emerging Topics in Computational Intelligence* **5**(2), 247–261 (2021). <https://doi.org/10.1109/TETCI.2019.2960474>
15. Karami, E., et al.: Adaptive Polar Active Contour for Segmentation and Tracking in Ultrasound Videos. *IEEE Trans. Circuits Syst. Video Technol.* **29**(4), 1209–1222 (2019). <https://doi.org/10.1109/TCSVT.2018.2818072>
16. Sharma, 2023, (December). Optimizing Knowledge Transfer in Sequential Models: Leveraging Residual Connections in Flow Transfer Learning for Lung Cancer Classification. In *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing* (pp. 1-8)
17. Ortiz-Gonzalez, A., et al.: Optical Flow-Guided Cine MRI Segmentation With Learned Corrections. *IEEE Trans. Med. Imaging* **43**(3), 940–953 (2024). <https://doi.org/10.1109/TMI.2023.3325766>
18. D. Morillo-Velepucha, et al. "Congestive heart failure prediction based on feature selection and machine learning algorithms," 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, 2022, pp. 1-6, <https://doi.org/10.23919/CISTI54924.2022.9820312>
19. Melillo, P., et al.: Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability. *IEEE J. Biomed. Health Inform.* **17**(3), 727–733 (2013). <https://doi.org/10.1109/JBHI.2013.2244902>
20. Kaiser, A., "Towards a method for early detection of congestive heart failure with an electrocardiogram and acoustic transducers," et al.: *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. New York, NY, USA **2012**, 1–5 (2012). <https://doi.org/10.1109/SPMB.2012.6469460>
21. Zhang, Y., "Congestive Heart Failure Detection Via Short-Time Electrocardiographic Monitoring For Fast Reference Advice In Urgent Medical Conditions," et al.: *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Honolulu, HI, USA **2018**, 2256–2259 (2018). <https://doi.org/10.1109/EMBC.2018.8512888>
22. Valenza, G., et al.: Mortality Prediction in Severe Congestive Heart Failure Patients With Multifractal Point-Process Modeling of Heartbeat Dynamics. *IEEE Trans. Biomed. Eng.* **65**(10), 2345–2354 (2018). <https://doi.org/10.1109/TBME.2018.2797158>
23. Mei-Yi Wu, et al., Radio-contrast medium exposure and dialysis risk in patients with chronic kidney disease and congestive heart failure: A case-only study, *International Journal of Cardiology*, Volume 324, 2021, Pages 199-204, ISSN 0167-5273, <https://doi.org/10.1016/j.ijcard.2020.09.014>
24. J. Zhang et al., "MLBF-Net: A Multi-Lead-Branch Fusion Network for Multi-Class Arrhythmia Classification Using 12-Lead ECG," in *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1-11, 2021, Art no. 1900211, <https://doi.org/10.1109/JTEHM.2021.3064675>

25. V. Jahmunah, et al. Computer-aided diagnosis of congestive heart failure using ECG signals - A review, *Physica Medica*, Volume 62, 2019, Pages 95-104, ISSN 1120-1797, <https://doi.org/10.1016/j.ejmp.2019.05.004>
26. Hou, B., et al.: LSTM-Based Auto-Encoder Model for ECG Arrhythmias Classification. *IEEE Trans. Instrum. Meas.* **69**(4), 1232–1240 (2020). <https://doi.org/10.1109/TIM.2019.2910342>
27. R.K. Tripathy, et al., Automated detection of congestive heart failure from electrocardiogram signal using Stockwell transform and hybrid classification scheme, *Computer Methods and Programs in Biomedicine*, Volume 173, 2019, Pages 53-65, ISSN 0169-2607, <https://doi.org/10.1016/j.cmpb.2019.03.008>
28. Guan, et al., "CobNet: Cross Attention on Object and Background for Few-Shot Segmentation." In 2022 26th International Conference on Pattern Recognition (ICPR), pp. 39-45. IEEE, 2022
29. De Marco, F., Finlay, D., Bond, R.R., "Classification of Premature Ventricular Contraction Using Deep Learning," *Computing in Cardiology*. Rimini, Italy **2020**, 1–4 (2020). <https://doi.org/10.22489/CinC.2020.311>
30. Mihaela Porumb, Ernesto Iadanza, Sebastiano Massaro, Leandro Pecchia, A convolutional neural network approach to detect congestive heart failure, *Biomedical Signal Processing and Control*, Volume 55, 2020, 101597, ISSN 1746-8094, <https://doi.org/10.1016/j.bspc.2019.101597>
31. De Marco F, Ferrucci F, Risi M, Tortora G (2022) Classification of QRS complexes to detect Premature Ventricular Contraction using machine learning techniques. *PLOS ONE* 17(8): e0268555. <https://doi.org/10.1371/journal.pone.0268555>



A Multimodal MRI-based Framework for Thyroid Cancer Diagnosis Using eXplainable Machine Learning

Ahmed Sharafeldeen¹, Hossam Magdy Balaha¹, Ali Mahmoud¹,
Reem Khaled², Saher Taman², Manar Mansour Hussein², Mohammed Ghazal³,
and Ayman El-Baz¹(✉)

¹ Bioengineering Department, J.B. Speed School of Engineering, University of Louisville, Louisville, KY, USA

aselba01@louisville.edu

² Department of Radiology, Faculty of Medicine, Mansoura University, Mansoura, Egypt

³ Electrical, Computer and Biomedical Engineering Department, Abu Dhabi University, Abu Dhabi, UAE

Abstract. Diagnosing thyroid cancer is notably challenging because of its diverse manifestations and the rising number of cases worldwide. Early detection and diagnosis of thyroid nodules' malignancy is crucial for reducing their progression. This paper introduces a novel computer-aided diagnosis (CAD) system that utilizes T2 and diffusion-weighted (DWI) magnetic resonance imaging (MRI) modalities to help diagnose thyroid cancer. First, the thyroid nodules are delineated from T2 and DWI modalities. Then, various features are extracted from these nodules, such as first order statistics (FOS), gray level co-occurrence matrix (GLCM), and gray level run length matrix (GLRLM), to capture texture and spatial information. To improve both the performance and interpretability of the model, outlier detection methods, such as the cluster-based local outlier factor (CBLOF), are utilized to identify deviations in the data. Finally, the extracted features from T2 and DWI modalities are fed into a multilayer perceptron (MLP) and LightGBM (LGBM) classifiers, respectively. Subsequently, the classifiers' outputs are integrated using a majority fusion approach for final diagnosis. The proposed system is evaluated on 55 thyroid nodule patients using a 10-folds cross-validation approach, achieving an accuracy of 99.48%. The reported results, based on integrating decisions from each MRI modality using a majority fusion approach, clearly demonstrate the effectiveness of the proposed framework compared to the performance of well-known pre-trained convolutional neural networks (CNNs).

Keywords: Computer Aided Diagnosis (CAD) · Machine Learning (ML) · SHapley Additive exPlanations (SHAP) · Thyroid Cancer Diagnosis

1 Introduction

Thyroid cancer is a common type of endocrine cancer that affects the thyroid gland, a gland shaped like a butterfly located in the neck's base. Diagnosing thyroid cancer is medically challenging due to its wide range of symptoms and varying levels of aggressiveness [4]. Thyroid cancer ranks as the 13th most common diagnosed cancer in the United States. The prevalence of thyroid cancer is estimated at approximately 2.2% of all cancer cases in the United States, with about 44,020 new cases reported in 2024 [19]. Despite its relatively low mortality rate compared to other cancers, early detection and precise diagnosis are essential for optimal treatment outcomes. The incidence of thyroid cancer has been rising steadily in recent decades, largely due to improved diagnostic methods that can identify smaller tumors and incidental findings during imaging examinations [6].

The main approaches for diagnosing thyroid cancer involve physical examination and imaging techniques (e.g., ultrasound, computed tomography [CT], and magnetic resonance imaging [MRI]), as well as fine-needle aspiration biopsy (FNAB). FNAB is considered the gold standard for evaluating thyroid nodules, where a fine needle is inserted into the thyroid gland to extract cellular material for analysis. Analysis of FNAB samples through histopathology determines if a thyroid nodule is benign or malignant, influencing subsequent treatment plans. Treatment options for thyroid cancer include surgical procedures (partial or total thyroidectomy), radioactive iodine therapy, and hormone replacement therapy [18].

The emergence of artificial intelligence (AI) and machine learning (ML) technologies has revolutionized the field of medical imaging and diagnostic pathology. AI and ML algorithms can assist radiologists and pathologists in interpreting imaging studies and FNAB specimens, improving diagnostic sensitivity and specificity of thyroid cancer. For example, Chaganti et al. [5] introduced a method that analyzed clinical data to predict Hashimoto's thyroiditis, autoimmune thyroiditis, binding protein, and non-thyroidal syndrome using machine learning and deep learning (DL) models. First, the authors utilized feature selection method, such as forward feature selection, backward feature elimination, bidirectional feature elimination, and ML-based feature selection (MLFS) using extra tree classifiers, to select the most relevant features. Then, they utilized different ML classifiers to investigate their performance. The latter were support vector machine (SVM), random forest, gradient boosting, AdaBoost, and logistic regression. Additionally, they assessed the effectiveness of DL methods, including long short-term memory (LSTM) network, convolutional neural network (CNN), and CNN-LSTM, in predicting thyroid diseases. Among the ML classifiers, random forest classifier yielded the highest accuracy of 99% when using features selected by MLFS method. In contrast, CNN outperformed other deep learning methods, achieving an accuracy of 93% with the original features. Wang et al. [20] conducted a study to investigate the predictive capability of ML-based multiparametric MRI radiomics in evaluating the aggressiveness of papillary thyroid carcinoma (PTC) prior to surgery. Thyroid nodules were manually delineated on MRI scans, followed by the extraction of 1393 radiomic features

from these nodules. The most relevant features were then identified using the least absolute shrinkage and selection operator (LASSO). Subsequently, a gradient boosting classifier was utilized to assess the aggressiveness of PTC, achieving an area under the receiver operating characteristic (ROC) curve (AUC) of 0.92. In [14], five different ML methods were utilized to differentiate between benign and malignant thyroid nodules in ultrasound images. The latter were SVM, deep neural network, center clustering method, logistic regression, and k-nearest neighbor. Their findings demonstrated that the deep neural network outperformed the other methods, achieving an accuracy of 87% in classifying the thyroid nodules. Another study [22] evaluated the predictive potential of MRI-based radiomics for identifying extrathyroidal extension (ETE) in PTC before surgery. Various radiomic features, including histogram, gray-level run length matrix (GLRLM), shape, gray-level size zone matrix (GLSZM), and gray-level co-occurrence matrix (GLCM), were extracted from the region of interest (ROI). Afterward, the most relevant features were selected using the maximum correlation minimum redundancy (mRMR) algorithm in combination with the LASSO. Finally, a radiomics predictive model was developed and tested on 132 patients, achieving an AUC of 0.87. A similar study [3] introduced a thyroid diagnostic system to classify thyroid nodules as benign or malignant by first extracting various radiomic features from T1, T2, and DWI-MRI images. The latter were wavelet transform, GLRLM, autoregressive model parameters, and GLCM. Afterward, feature selection approaches, including mutual information, Fisher coefficient, and classification error probability and average correlation coefficients, were utilized to identify the most relevant features. Finally, a linear discriminant analysis (LDA) classifier was employed to diagnose thyroid nodules. Zhang et al. [24] proposed a multi-channel CNN thyroid diagnostic system utilizing multiparametric MRI scans. First, three-channels features were extracted from T1, T2, and contrast-enhanced T1 MRI images using two-layers CNN. Then, a multi-feature association layer was introduced to integrate the features extracted from these three models. Finally, a fully connected layer was employed and fed with features extracted from the multi-feature association layer to classify thyroid regions as normal, benign, or malignant. A recent study [23] examined the precision of various well-known pretrained CNNs, such as ResNet50, MobileNet, AlexNet, ShuffleNet, and NasNetMobile, in automatically extracting features for identifying the most effective descriptor of thyroid nodules in ultrasound images. The authors employed an SVM classifier to classify the features extracted from each CNN as benign or malignant. The study found that ResNet50 was the optimal model for extracting relevant features to distinguish between benign and malignant nodules.

The scarcity of research on MRI-based thyroid classification motivated us to create an innovative computer-aided diagnosis (CAD) system. This system leverages texture features extracted from segmented thyroid nodules to distinguish between benign and malignant nodules using T2 and diffusion-weighted (DWI)-MRI, with the following contributions: 1) We investigate the ability of multimodal MRI in distinguishing between benign and malignant thyroid nod-

ules. 2) Different radiomics features are employed to distinguish between these thyroid nodules. 3) We investigate the impact of outlier detection methods on the performance of ML classifiers in distinguishing between benign and malignant nodules by removing unusual instances from the training dataset. 4) SHapley Additive exPlanations (SHAP) is used to visualize and interpret the model predictions and feature importance. 5) The proposed system outperforms various well-known pretrained CNNs, highlighting its potential to be used as an aided diagnostic tool in clinical treatment.

2 Materials and Methods

The current study presents a CAD framework, depicted in Fig. 1, for diagnosing thyroid cancer using T2 and DWI modalities. It consists of several phases, including: 1) Different features are extracted from thyroid nodules in T2 and DWI modalities, delineated manually by a radiologist. 2) An outlier detection method is employed to remove anomalous data points extracted from the DWI modality. 3) Features extracted from each modality are fed individually into an ML classifier. Then, a majority voting approach is used to obtain the final diagnosis. 4) SHAP AI explainability is utilized to visualize the importance of features in the model output.

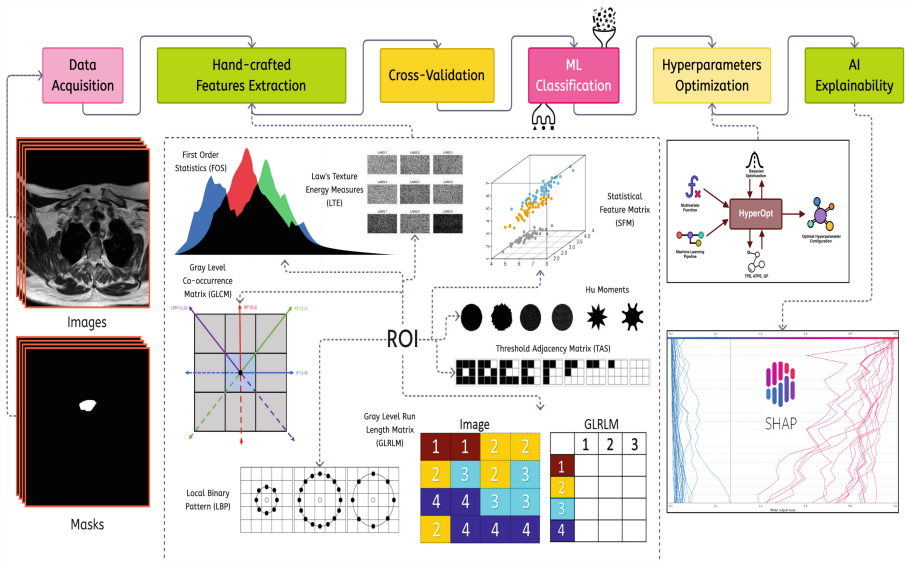


Fig. 1. The suggested Computer Aided Diagnosis (CAD) framework thyroid cancer diagnosis from T2 and DWI modalities.

2.1 Materials

The thyroid gland was examined using a 1.5 T Ingenia MR scanner from Philips Medical Systems in the Netherlands. A specialized surface coil designed for the head and neck was utilized for imaging. T2-weighted images were obtained swiftly with a pulse sequence featuring a TR (repetition time) of 6000 ms and a TE (echo time) of 80 ms. The resulting volumes consisted of axial slices with a thickness of 5 mm, interspersed with either a 1 mm or 2 mm gap between slices. The field of view (FOV) varied between 20 cm and 25 cm, and data was sampled on a 256×256 acquisition matrix. For diffusion-weighted imaging (DWI) of the same volume, a specialized sequence was employed, characterized by a rapid acquisition process using a single-shot, spin-echo, echo-planar imaging technique. The parameters included a TR of 10,000 ms, TE of 108 ms, and a bandwidth of 125 kHz. The axial slices for diffusion-weighted imaging were also 5 mm thick, with a 1 mm gap between slices. The FOV ranged from 25 cm to 30 cm, and the acquisition matrix was 256×256 . To capture diffusion in multiple directions, gradients were applied in three orthogonal directions. Two different b-factors were utilized, specifically $b = 500 \text{ s/mm}^2$ and $b = 1,000 \text{ s/mm}^2$, along with a $b = 0$ scan, resulting in a total of seven acquisitions for each slice location. This retrospective study investigated the MRI characteristics of 55 patients diagnosed with thyroid nodules confirmed through pathology, ranging in size from 1 to 3 cm. Among them, malignant nodules were identified in 20 patients, while benign nodules were observed in 35 patients. Samples from the utilized thyroid cancer dataset from the two modalities (T2 and DWI) are presented in Fig. 2.

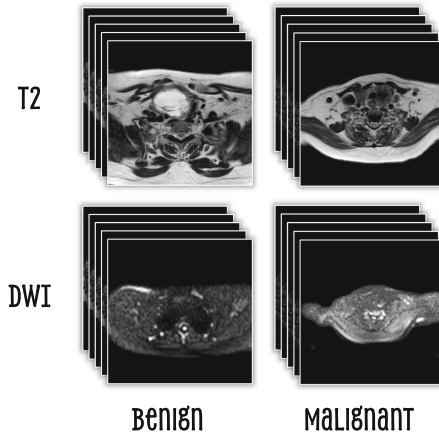


Fig. 2. Samples from the utilized thyroid cancer dataset from the two modalities (T2 and DWI).

2.2 Feature Extraction from T2 and DWI Modalities

The current study investigates the extraction of various features from T2 and DWI modalities to provide a comprehensive diagnosis of thyroid conditions. Each feature extraction technique is based on specific hypotheses aimed at capturing distinct characteristics of thyroid tissue morphology, texture, and spatial distribution.

First Order Statistics (FOS): FOS captures essential information about tissue composition and density variations in the thyroid region by calculating statistical measures such as mean, standard deviation, and entropy from pixel intensity distributions.

Gray Level Co-occurrence Matrix (GLCM): GLCM analysis in T2 and DWI images seeks to extract texture information by examining the spatial relationships between pixel intensities. This method quantifies the occurrence of pixel pairs with particular intensity values and spatial configurations. Features, such as contrast and entropy, derived from GLCM, provide valuable information about the texture complexity and variability found in thyroid tissue.

Gray Level Run Length Matrix (GLRLM): Similar to GLCM, GLRLM extraction in T2 and DWI modalities seeks to reveal patterns and structures within the thyroid region by analyzing the distribution of gray level runs. Features, such as short run emphasis and long run emphasis, offer information about the length and intensity of homogeneous regions, aiding in assessing thyroid homogeneity and spatial organization.

Local Binary Pattern (LBP): The LBP method in T2 and DWI images seeks to capture distinctive textural details that remain consistent despite changes in lighting conditions. By encoding pixel relationships into binary patterns and analyzing their distribution, LBP features highlight textural patterns and spatial variations within the thyroid tissue, enabling robust texture characterization.

Threshold Adjacency Matrix (TAS): TAS extraction in T2 and DWI modalities analyzes the relationships between adjacent pixels at various intensity thresholds to encode significant spatial details. By examining connectivity patterns between thresholded regions, TAS features aim to capture the spatial arrangement and structural complexity of the thyroid tissue, aiding in its characterization.

Hu Moments: Extracting Hu moments from T2 and DWI images leverages certain image moments' invariant properties to translation, scale, and rotation. These moments extracted from the thyroid region serve as concise descriptors of its shape and spatial distribution, enhancing geometric characterization and distinguishing it from surrounding tissue.

Statistical Feature Matrix (SFM): SFM extraction in T2 and DWI modalities examines the statistical properties of pixel pairs across different distances to encode details about image texture and spatial organization. By analyzing how pixel pairs co-occur at various spatial scales, SFM features effectively capture textural patterns, periodicity, and roughness present within the thyroid tissue.

Law's Texture Energy Measures (LTE): LTE extraction in T2 and DWI images entails deriving texture features at various spatial scales by convolving

simple masks with the image. The distribution of energy across different texture components reflects the composition and organization of tissue within the thyroid region, facilitating a thorough characterization of its texture.

Shape Parameters: Extracting shape parameters from T2 and DWI images offers valuable insights into the morphology and spatial extent of the thyroid. These parameters, including maximum length, area, and perimeter, quantify aspects of thyroid shape, size, and boundary irregularity. This analysis aids in morphological assessment and classification of thyroid characteristics.

2.3 Outliers Detection

To improve the performance of the training process, detecting unusual instances (i.e., outliers) is crucial. Removing these outliers from the training process is essential for improving the performance and interpretability of ML models used for thyroid cancer diagnosis across various modalities and their fusion. The outlier detection methods include angle-based outlier detector (ABOD) [11], histogram-based outlier detection (HBOS) [7], cluster-based local outlier factor (CBLOF), k-nearest neighbors (KNN), isolation forest (IForest) [13], outlier detection with kernel density estimation (KDE) [12], isolation-based anomaly detection using nearest-neighbor ensembles (INNE) [2], Gaussian mixture model (GMM) outlier detection, and minimum covariance determinant (MCD) [8]. ABOD effectively identifies outliers by analyzing the unusual angles formed by data points in the feature space. HBOS assesses the likelihood of outliers by analyzing feature value histograms. CBLOF detects outliers by identifying instances with sparse cluster assignments or clusters with low density, while IForest divides the feature space to effectively isolate outliers. KDE estimates density functions to identify outliers occurring in low-density regions. INNE constructs ensembles of nearest neighbors to isolate consistent outliers. GMM detects outliers by creating a probabilistic model based on multiple Gaussian distributions. This approach effectively demonstrates its capability to model complex data distributions. Moreover, MCD provides a robust estimation of data distribution covariance despite the presence of outliers. Choosing suitable outlier detection method helps the classifier find the appropriate boundary to separate benign and malignant nodules. Based on the experimental results, CBLOF was identified as the most effective outlier detection method for the ML classifier using features extracted from the DWI modality. Meanwhile, the ML classifier achieved optimal results when fed with features extracted from the T2 modality.

2.4 Thyroid Diagnosis Using T2 and DWI Modalities

In this paper, T2 and DWI modalities are employed to diagnose thyroid conditions using machine learning. Each modality enables a comprehensive evaluation of thyroid gland by offering distinctive perspectives on the structure and function of the thyroid gland. The aim is to identify the most effective diagnostic approach that maximizes accuracy and reliability in detecting thyroid abnormalities

using T2 and DWI imaging techniques. We utilize different classification algorithms including logistic regression (LR), KNN, decision trees (DT), random forest (RF), AdaBoost, XGBoost (XGB), gradient boosting (GB), LightGBM (LGBM), SVM, extra trees (ET), and multilayer perceptron (MLP). Each algorithm offers specific strengths tailored to handle the complexities of thyroid diagnosis.

To enhance the algorithms' performance, we utilize the tree-structured parzen estimator (TPE) for hyperparameter optimization. TPE directs the exploration of hyperparameter configurations using Bayesian optimization techniques, effectively exploring the hyperparameter space. In contrast to traditional grid or random search methods, TPE prioritizes promising areas in the hyperparameter space, thereby boosting the performance of classification algorithms in thyroid diagnosis [21].

2.5 Explanation of SHapley Additive exPlanations (SHAP)

To explain the outcomes of ML models by assigning importance scores to input features, eXplainable AI [1,9], specifically SHapley Additive exPlanations (SHAP) framework, is adopted in the system pipeline. SHAP framework leverages Shapley values derived from cooperative game theory, measuring the contribution of each feature to the model's prediction [9]. The fundamental idea behind SHAP is the Shapley value formula, which calculates the marginal contribution of each feature to the prediction. The Shapley value for feature i ($\phi_i(v)$) is determined by considering all possible subsets of features and their respective contributions to the prediction [10], which is defined as follows:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \times [f(S \cup \{i\}) - f(S)] \quad (1)$$

where N denotes the set of all features, S signifies a subset of features excluding i , $f(S \cup \{i\})$ represents the model's prediction when considering the features in S along with feature i , and $f(S)$ is the model's prediction when considering only the features in S .

3 Experiments and Discussion

The current study is conducted within a Python-based software environment, running on Windows 11 and utilizing Anaconda as the preferred distribution platform. The hardware configuration comprises an NVIDIA GPU with 6GB of memory, 256GB of RAM, and an Intel Core i7 processor. To assess the effectiveness of our predictive models in diagnosing thyroid conditions, a comprehensive evaluation framework is employed. This framework incorporates various evaluation metrics such as accuracy, precision, recall, specificity, F1 score, Intersection over Union (IoU), Balanced Accuracy (BAC), Matthews Correlation Coefficient (MCC), Youden's J statistic, and Yule's Q statistic, ensuring a thorough assessment of model performance.

To show the effectiveness of various ML algorithms applied to different modalities, namely T2 and DWI, as well as their fusion, their performance is evaluated, as demonstrated in Table 1. As presented in the table, the MLP classifier outperformed other classifiers when fed with features extracted from T2 modality, achieving the highest accuracy of 93.91% among them. Its robustness is also noteworthy, as it maintains high precision, recall, and specificity scores, resulting in an impressive F1 score of 95.71%. This suggests that MLP is well-suited for distinguishing between different tissue types or abnormalities within T2-weighted images. On the other hand, for the DWI modality, the LGBM classifier demonstrates superior performance, with an accuracy of 97.42% and a balanced combination of precision, recall, and specificity, resulting in an F1 score of 98.26%. This indicates that LGBM is adept at detecting subtle abnormalities or lesions within diffusion-weighted images, making it a promising choice for diagnostic purposes in clinical settings. When considering the fusion of modalities using majority voting, both the intersection (i.e., the common cases retaining both T2 and DWI-based features after outlier detection) and union (i.e., integration of both T2 and DWI-based classification) approaches yield significantly improved performance compared to individual modalities alone. The fusion models achieve near-perfect accuracy, precision, recall, and specificity, indicating a synergistic effect where the complementary information from T2 and DWI enhances the overall diagnostic capability. This underscores the importance of multimodal imaging in improving diagnostic accuracy and reliability.

Table 1. Tabular presentation of performance results for various ML algorithms applied to the T2 and DWI modalities using 10-fold cross-validation. Bold highlights the best classifier for each modality.

Modality	Classifier	Scaler	Outlier	Records	Accuracy	Precision	Recall	Specificity	F1	IoU	BAC	MCC	Youden	Yule	
T2	AdaBoost	Robust	CBLOF	259	88.42	90.26	94.12	73.61	92.15	85.44	83.86	70.35	67.73	95.62	
	CatBoost	MinMax	INNE	259	91.51	91.04	97.86	75.00	94.33	89.27	86.43	78.30	72.86	98.55	
	DT	STD	INNE	259	81.47	90.12	83.33	76.71	86.59	76.35	80.02	57.20	60.05	88.55	
	ET	Robust	INNE	259	93.44	92.46	98.92	79.45	95.58	91.54	89.19	83.58	78.38	99.44	
	GB	None	INNE	259	87.64	91.01	91.98	76.39	91.49	84.31	84.18	68.97	68.37	94.75	
	HGB	None	INNE	259	89.96	91.15	95.11	77.33	93.09	87.06	86.22	75.03	72.44	97.03	
	KNN	Robust	ABOD	264	88.26	88.67	95.74	69.74	92.07	85.31	82.74	70.34	65.48	96.22	
	LGBM	MaxAbs	ABOD	264	90.53	91.79	95.21	78.95	93.47	87.75	87.08	76.42	74.16	97.35	
	LR	Robust	KDE	345	85.51	88.14	91.77	70.59	89.92	81.68	81.18	64.35	62.36	92.80	
	MLP	STD	None	345	93.91	95.12	96.30	88.24	95.71	91.76	92.27	85.28	84.53	98.98	
	RF	None	KNN	292	88.36	88.09	97.18	64.56	92.41	85.89	80.87	69.20	61.74	96.87	
	SVM	STD	IForest	259	86.87	85.96	98.99	47.54	92.02	85.22	73.27	60.83	46.53	97.77	
	XGB	None	ABOD	264	89.39	89.60	96.28	72.37	92.82	86.60	84.32	73.32	68.65	97.09	
	DWI	AdaBoost	None	CBLOF	233	96.57	96.57	98.83	90.32	97.69	95.48	94.58	91.12	89.15	99.75
		CatBoost	Robust	GMM	233	95.71	96.02	98.26	88.52	97.13	94.41	93.39	88.75	86.78	99.54
DT		None	None	311	93.89	95.52	95.95	88.76	95.73	91.81	92.35	85.00	84.71	98.94	
ET		None	ABOD	234	96.58	96.02	99.41	89.06	97.69	95.48	94.24	91.34	88.47	99.85	
GB		MaxAbs	KDE	311	95.82	95.63	98.65	88.76	97.12	94.40	93.71	89.66	87.41	99.65	
HGB		MinMax	KNN	247	95.14	96.61	96.61	91.43	96.61	93.44	94.02	88.04	88.04	99.34	
KNN		MinMax	HBOS	233	92.70	92.17	97.45	82.89	94.74	90.00	90.17	83.22	80.35	98.93	
LGBM		STD	CBLOF	233	97.42	97.69	98.83	93.55	98.26	96.57	96.19	93.36	92.38	99.84	
LR		STD	MCD	233	90.56	92.90	94.01	81.82	93.45	87.71	87.92	76.55	75.83	97.21	
MLP		STD	GMM	233	94.42	95.43	97.09	86.89	96.25	92.78	91.99	85.38	83.98	99.10	
RF		Robust	KNN	247	95.14	95.58	97.74	88.57	96.65	93.51	93.16	87.90	86.31	99.41	
SVM		STD	HBOS	233	88.84	87.01	98.09	69.74	92.22	85.56	83.91	74.42	67.83	98.32	
XGB		STD	KNN	247	96.76	96.69	98.87	91.43	97.77	95.63	95.15	91.96	90.30	99.79	
Fusion (Intersection)					110	99.09	98.80	100	96.43	99.39	98.80	98.21	97.60	96.43	100
Fusion (Union)					383	99.48	99.27	100	98.20	99.63	99.27	99.10	98.73	98.20	100

Fig. 3 presents the ROC curves for the two modalities across the different categories, emphasizing the proposed system’s capability to discriminate between benign and malignant thyroid tumors. Figs. 4 and 5 present detailed visual SHAP explanation, focusing on 50 random samples and highlighting the top 20 features through a decision plot. This plot illustrates how each feature contributes to the model’s output for a specific instance, offering valuable insights into the model’s behavior and feature importance. In contrast, Fig. 6 provides a comprehensive breakdown of feature contributions, with each modality (T2 and DWI) represented by the same sample. Features are displayed horizontally, with arrows indicating both the direction and magnitude of impact for each feature. Together, these visualizations offer a granular understanding of the model’s decision-making process, aiding interpretation, debugging, and validation efforts.

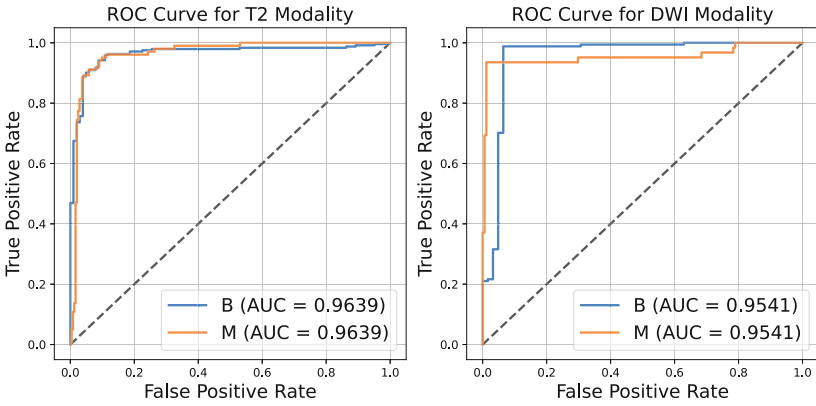


Fig. 3. Visual presentation of the ROC curves for the two modalities across the different categories.

3.1 Comparison with Deep Learning

An experiment was conducted using pretrained convolutional neural networks (CNNs), including ResNet50V2, MobileNet, MobileNetV2, InceptionResNetV2, DenseNet169, Xception, NASNetLarge, ResNet152V2, ResNet50, DenseNet201, ResNet101V2, DenseNet121, InceptionV3, VGG16, VGG19, and NASNetMobile, to explore whether these models could outperform the proposed fusion approach. Table 2 presents the performance results for the two modalities after conducting experiments with pretrained CNNs using 80%:20% train-test-split ratio. Overall, these models achieved accuracy scores ranging from 70% to 79%, with BAC scores between 50% and 64%, and precision scores between 70% and 77%. It is important to highlight that the recall scores were consistently at 100%, indicating that all positive instances were correctly identified. Conversely, the specificity scores were generally low, often at 0%, implying that the models struggled to

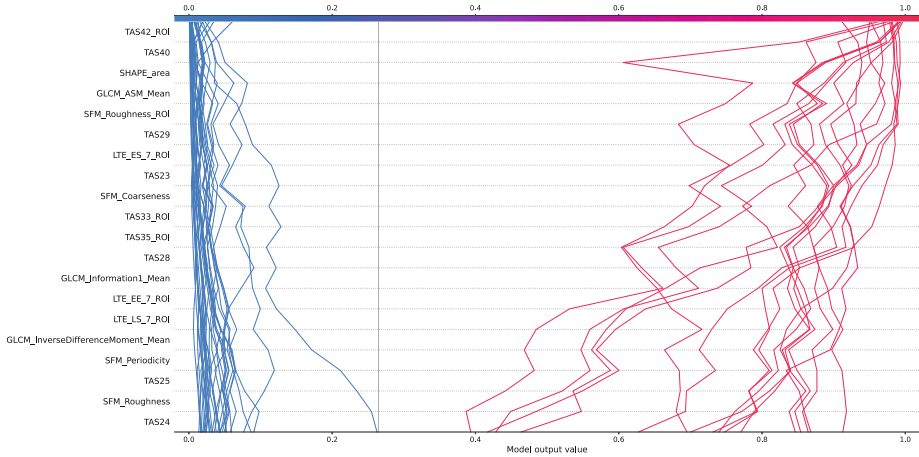


Fig. 4. Visual SHAP explanation displaying 50 random samples, emphasizing the top 20 features for T2 using a decision plot to elucidate each feature’s contribution to the model’s output for a specific instance. Features are arranged along the y-axis, with their SHAP values depicted by bars.

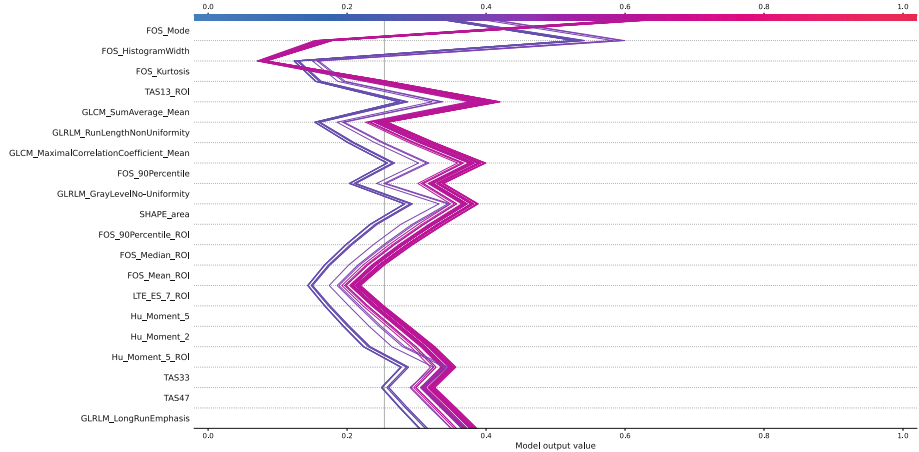


Fig. 5. Visual SHAP explanation displaying 50 random samples, emphasizing the top 20 features for DWI using a decision plot to elucidate each feature’s contribution to the model’s output for a specific instance. Features are arranged along the y-axis, with their SHAP values depicted by bars.

correctly identify negative instances. This issue may be attributed due to the imbalance in the dataset, the small dataset size, or the low resolution of DWI-MRI images. It appears that the performance of the pretrained CNN models, as shown in the table, may not be suitable when compared with ML approaches and hand-crafted features.

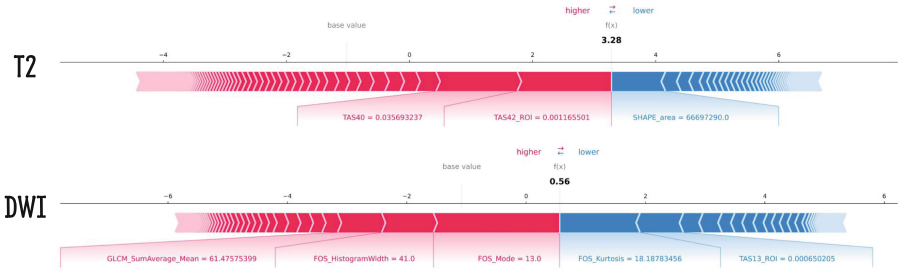


Fig. 6. Visual SHAP explanation showcasing a sample from each modality of the two modalities (T2 and DWI), emphasizing the detailed breakdown of feature contributions. Features are displayed horizontally, with arrows indicating both direction and magnitude of impact for each feature.

Table 2. Tabular presentation of performance results for the fusion of the two modalities after conducting experiments with pretrained CNNs using 80%:20% train-test-split ratio.

T2 Modality						DWI Modality							
Model	Accuracy	BAC	Precision	Recall	Specificity	F1	Model	Accuracy	BAC	Precision	Recall	Specificity	F1
MobileNet	79.74	64.61	77.89	100	29.21	87.57	MobileNet	70.43	50.00	70.43	100	0	82.65
MobileNetV2	74.60	55.95	73.91	99.55	12.36	84.84	MobileNetV2	70.43	50.00	70.43	100	0	82.65
DenseNet121	71.70	50.56	71.61	100	1.12	83.46	DenseNet121	70.43	50.00	70.43	100	0	82.65
DenseNet169	76.53	58.99	75.25	100	17.98	85.88	DenseNet169	70.43	50.00	70.43	100	0	82.65
DenseNet201	71.38	50.00	71.38	100	0	83.30	DenseNet201	70.43	50.00	70.43	100	0	82.65
VGG16	71.38	50.00	71.38	100	0	83.30	VGG16	70.43	50.00	70.43	100	0	82.65
VGG19	71.38	50.00	71.38	100	0	83.30	VGG19	70.43	50.00	70.43	100	0	82.65
ResNet50	71.38	50.00	71.38	100	0	83.30	ResNet50	70.43	50.00	70.43	100	0	82.65
ResNet101	71.38	50.00	71.38	100	0	83.30	ResNet101	70.43	50.00	70.43	100	0	82.65
ResNet152	71.38	50.00	71.38	100	0	83.30	ResNet152	70.43	50.00	70.43	100	0	82.65
ResNet50V2	76.53	58.99	75.25	100	17.98	85.88	ResNet50V2	70.43	50.00	70.43	100	0	82.65
ResNet101V2	78.78	62.92	77.08	100	25.84	87.06	ResNet101V2	70.43	50.00	70.43	100	0	82.65
ResNet152V2	72.99	52.81	72.55	100	5.62	84.09	ResNet152V2	70.43	50.00	70.43	100	0	82.65
InceptionV3	71.38	50.00	71.38	100	0	83.30	InceptionV3	70.43	50.00	70.43	100	0	82.65
InceptionResNetV2	77.17	61.46	76.49	98.20	24.72	86.00	InceptionResNetV2	70.43	50.00	70.43	100	0	82.65
NASNetLarge	71.38	50.00	71.38	100	0	83.30	NASNetLarge	70.43	50.00	70.43	100	0	82.65
NASNetMobile	75.24	56.74	74.25	100	13.48	85.22	NASNetMobile	70.43	50.00	70.43	100	0	82.65
Xception	71.38	50.00	71.38	100	0	83.30	Xception	70.43	50.00	70.43	100	0	82.65
EfficientNetB0	71.38	50.00	71.38	100	0	83.30	EfficientNetB0	70.43	50.00	70.43	100	0	82.65
EfficientNetB1	71.38	50.00	71.38	100	0	83.30	EfficientNetB1	70.43	50.00	70.43	100	0	82.65
EfficientNetB2	71.38	50.00	71.38	100	0	83.30	EfficientNetB2	70.43	50.00	70.43	100	0	82.65
EfficientNetB3	71.38	50.00	71.38	100	0	83.30	EfficientNetB3	70.43	50.00	70.43	100	0	82.65

To further demonstrate the capability of the proposed diagnostic system, its performance is compared with that of our previous works, as depicted in Table 3. As shown in the table, the performance of the proposed system is significantly improved by 12% compared to the lowest-performing system, underscoring its effectiveness in detecting the malignancy of thyroid tumors.

Table 3. A tabular presentation comparing the performance results of previous related works with those of the proposed system.

Model	Accuracy (%)
Naglah et. al [15]	87
Sharafeldeen [16]	93.65
Sharafeldeen [17]	95.5
Proposed system	99.48

4 Conclusions and Future Directions

In conclusion, this study introduces a CAD system for diagnosing thyroid cancer, utilizing T2 and DWI MRI modalities. By leveraging diverse feature extraction methods and outlier detection techniques, the CAD framework offers a robust approach to capturing intricate tissue features and identifying abnormalities within the thyroid gland. The conducted experiments highlight the superior performance of the fusion models incorporating T2 and DWI modalities compared to individual modality-based classification as well as pretrained convolutional neural networks (CNNs). The fusion models exhibit near-perfect accuracy and comprehensive evaluation metrics, underscoring the synergistic effect of multi-modal imaging in enhancing diagnostic capabilities. In the future, we aim to integrate additional imaging modalities or biomarkers to enhance the accuracy of thyroid cancer diagnosis, as well as to investigate the performance of proposed system on larger datasets. Moreover, exploring molecular imaging techniques or genetic markers may provide valuable insights into tumor biology and prognosis, complementing structural information from T2 and DW MRI modalities.

References

1. Aljadani, A., Alharthi, B., Farsi, M.A., Balaha, H.M., Badawy, M., Elhosseini, M.A.: Mathematical modeling and analysis of credit scoring using the lime explainer: A comprehensive approach. *Mathematics* **11**(19), 4055 (2023)
2. Bandaragoda, T.R., Ting, K.M., Albrecht, D., Liu, F.T., Zhu, Y., Wells, J.R.: Isolation-based anomaly detection using nearest-neighbor ensembles. *Comput. Intell.* **34**(4), 968–998 (2018)
3. Brown, A.M., Nagala, S., McLean, M.A., Lu, Y., Scoffings, D., Apte, A., Gonen, M., Stambuk, H.E., Shaha, A.R., Tuttle, R.M., Deasy, J.O., Priest, A.N., Jani, P., Shukla-Dave, A., Griffiths, J.: Multi-institutional validation of a novel textural analysis tool for preoperative stratification of suspected thyroid tumors on diffusion-weighted mri. *Magn. Reson. Med.* **75**(4), 1708–1716 (2015). <https://doi.org/10.1002/mrm.25743>
4. Cabanillas, M.E., McFadden, D.G., Durante, C.: Thyroid cancer. *The Lancet* **388**(10061), 2783–2795 (2016)
5. Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J.L.V., Rodríguez, C.L., Ashraf, I.: Thyroid disease prediction using selective features and machine

- learning techniques. *Cancers* **14**(16), 3914 (Aug 2022). <https://doi.org/10.3390/cancers14163914>
6. Davies, L., Welch, H.G.: Current thyroid cancer trends in the united states. *JAMA otolaryngology-head & neck surgery* **140**(4), 317–322 (2014)
 7. Goldstein, M., Dengel, A.: Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track* **1**, 59–63 (2012)
 8. Hardin, J., Rocke, D.M.: Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis* **44**(4), 625–638 (2004)
 9. Huang, Y., Wang, X., Cao, Y., Li, M., Li, L., Chen, H., Tang, S., Lan, X., Jiang, F., Zhang, J.: Multiparametric mri model to predict molecular subtypes of breast cancer using shapley additive explanations interpretability analysis. *Diagnostic and Interventional Imaging* (2024)
 10. Kim, Y., Kim, Y.: Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. *Sustain. Urban Areas* **79**, 103677 (2022)
 11. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 444–452 (2008)
 12. Latecki, L.J., Lazarevic, A., Pokrajac, D.: Outlier detection with kernel density functions. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. pp. 61–75. Springer (2007)
 13. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: *2008 eighth ieee international conference on data mining*. pp. 413–422. IEEE (2008)
 14. Ma, X., Xi, B., Zhang, Y., Zhu, L., Sui, X., Tian, G., Yang, J.: A machine learning-based diagnosis of thyroid cancer using thyroid nodules ultrasound images. *Curr. Bioinform.* **15**(4), 349–358 (2020). <https://doi.org/10.2174/1574893614666191017091959>
 15. Naglah, A., Khalifa, F., Khaled, R., Abdel Razek, A.A.K., Ghazal, M., Giridharan, G., El-Baz, A.: Novel mri-based cad system for early detection of thyroid cancer using multi-input cnn. *Sensors* **21**(11), 3878 (Jun 2021) <https://doi.org/10.3390/s21113878>
 16. Sharafeldeen, A., Elsharkawy, M., Shaffie, A., Khalifa, F., Soliman, A., Naglah, A., Khaled, R., Hussein, M.M., Alrahmawy, M., Elmougy, S., Yousaf, J., Ghazal, M., El-Baz, A.: Thyroid cancer diagnostic system using magnetic resonance imaging. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE (Aug 2022) <https://doi.org/10.1109/icpr56361.2022.9956125>
 17. Sharafeldeen, A., Elsharkawy, M., Khaled, R., Shaffie, A., Khalifa, F., Soliman, A., Abdel Razek, A.A.k., Hussein, M.M., Taman, S., Naglah, A., Alrahmawy, M., Elmougy, S., Yousaf, J., Ghazal, M., El-Baz, A.: Texture and shape analysis of diffusion-weighted imaging for thyroid nodules classification using machine learning. *Medical Physics* **49**(2), 988–999 (Dec 2021) <https://doi.org/10.1002/mp.15399>
 18. Sharifovna, Y.H.: Thyroid cancer diagnostics, classification, staging. *Ijtimoiy fanlarda innovasiya onlayn ilmiy jurnali* **1**(5), 63–69 (2021)
 19. Siegel, R.L., Giaquinto, A.N., Jemal, A.: *Cancer statistics, 2024*. CA: A Cancer Journal for Clinicians **74**(1), 12–49 (Jan 2024). <https://doi.org/10.3322/caac.21820>
 20. Wang, H., Song, B., Ye, N., Ren, J., Sun, X., Dai, Z., Zhang, Y., Chen, B.T.: Machine learning-based multiparametric mri radiomics for predicting the aggressiveness of papillary thyroid carcinoma. *European Journal of Radiology* **122**, 108755 (Jan 2020) <https://doi.org/10.1016/j.ejrad.2019.108755>

21. Watanabe, S.: Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv preprint [arXiv:2304.11127](https://arxiv.org/abs/2304.11127) (2023)
22. Wei, R., Wang, H., Wang, L., Hu, W., Sun, X., Dai, Z., Zhu, J., Li, H., Ge, Y., Song, B.: Radiomics based on multiparametric mri for extrathyroidal extension feature prediction in papillary thyroid cancer. *BMC Medical Imaging* **21**(1) (Feb 2021) <https://doi.org/10.1186/s12880-021-00553-z>
23. Yadav, N., Dass, R., Virmani, J.: Deep learning-based cad system design for thyroid tumor characterization using ultrasound images. *Multimedia Tools and Applications* **83**(14), 43071–43113 (2023). <https://doi.org/10.1007/s11042-023-17137-4>
24. Zhang, R., Liu, Q., Cui, H., Wang, X., Song, S., Huang, G., Feng, D.: Thyroid classification via new multi-channel feature association and learning from multi-modality mri images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (Apr 2018). <https://doi.org/10.1109/isbi.2018.8363573>

Author Index

A

Abdelrahim, Mostafa 250
Agarwal, Kushagra 105
Ali, Mahmoud 250
Alksas, Ahmed 392
Ameer, P. M. 376
Anand, Saket 56
Augenstein, Christoph 359

B

Bai, Jinfeng 192
Balaha, Hossam Magdy 250, 438
Basioti, Kalliopi 39
Bauer, Markus 359
Beaudett, Benjamin 56
Bollu, Tharun Kumar Reddy 296

C

Cao, Junhao 89
Chatterjee, Swarnadip 408
Chen, Junyi 89
Chen, Xilin 192
Contractor, Sohail 392

D

Desai, Nandakishor 73
Dey, Sayantan 281
Dhar, Joy 311

E

El-Baz, Ayman 250, 392, 438
Elkhouly, Abdelrhman 392

F

Franczyk, Bogdan 359

G

Garcia-Zapirain, Begonya 250
George, Sudhish N. 376
Ghazal, Mohammed 392, 438

Göksel, Orcun 408
Goyal, Puneet 311
Gruhl, Christian 1
Gurwin, Adam 359
Gyanchandani, Bhavesh 422

H

He, Yujiang 1
Huang, Sib0 89
Huang, Zhixin 1
Hussein, Manar Mansour 438

I

Iwana, Brian Kenji 160

J

Ji, Zhilong 192
Jiang, Feng 219
Joseph, Nirmal 376

K

Kadam, Gargi 266
Kalbande, Dhananjay R. 266
Kan, Meina 192
Karawia, Abdelrahman 392
Khaled, Reem 438
Khudri, Mohmaed 250
Kolekar, Maheshkumar H. 266
Kumar, Santosh 422

L

Lee, Jiseok 160
Li, Jiatong 39
Li, Kuan 344
Li, Stan Z. 22
Li, Yanyi 344
Liang, Shenyuan 56
Lindblad, Joakim 408
Luo, Guibo 176

M

Mahmoud, Ali 438
Malkiewicz, Bartosz 359

N

Nivarthi, Chandana Priya 1

O

Osa-Sanchez, Ainhoa 250
Oza, Aditya 422

P

Palaniswami, Marimuthu 73
Pandey, Shraddha 105
Parekkattil, Adarsh V. 296
Patel, Sanskriti 422
Patil, Hemant A. 234
Patil, Shanta Hardas 146
Pavlovic, Vladimir 39
Ponomarev, Andrew 120

R

Raja, Kiran 376
Rana, Kapil 311
Roy, Abhinav 422
Roy, Partha Pratim 281

S

Saleh, Gehad A. 392
Shah, Arth J. 234
Shan, Shiguang 192
Sharafeldeen, Ahmed 438
Shehata, Mohamed 392
Sick, Bernhard 1
Siddhad, Gourav 281
Singh, Anand Vir 105
Singh, Maneet 105
Sladoje, Nataša 408
Somavarapu, Tarun 105
Sonawane, Yash 266

Srivastava, Anuj 56
Suthar, Manish 234

T

Taman, Saher 438
Tiwarekar, Sanika 266
Turaga, Pavan 56

V

Varun, Sanjeev Kumar 296
Verma, Shantanu 105

W

Wang, Guangshuo 176
Wang, Jianjia 131
Wang, Shiyun 328
Wang, Yi 219
Wen, Jing 219
Wu, Xing 131

X

Xiao, Xiongjiang 344
Xu, Pin 344
Xu, Yongchao 328
Xu, Yuanyuan 192

Y

Yadav, Agnesh Chandra 266
Yang, Zhiyuan 208
Yao, Yuchong 73
Yeo, Si Yong 208
Yin, Jianping 344
Yuen, Weimin 344

Z

Zeng, Zhiqiang 208
Zhang, Bo 208
Zhang, Dongyu 89
Zhao, Xunda 131
Zheng, Jiangbin 22
Zhu, Yuesheng 176