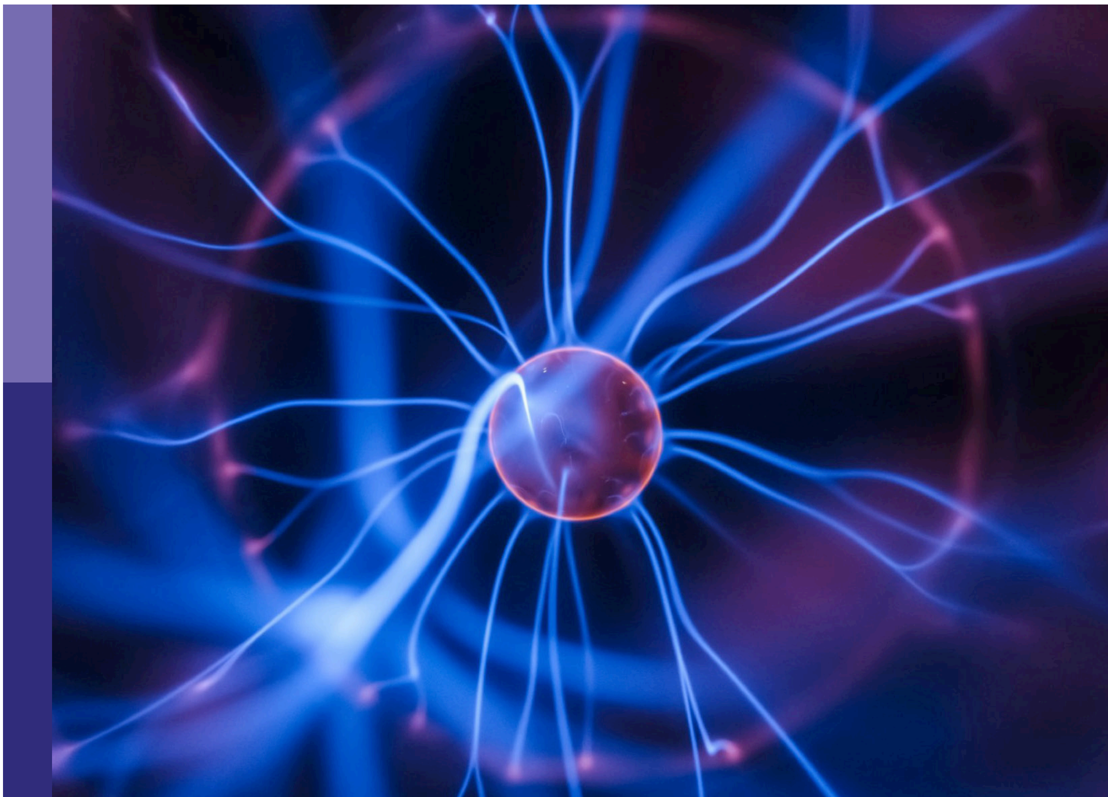
Research Topics

Multi-sensor imaging and fusion: Methods, evaluations, and applications, volume II

Edited by
Zhiqin Zhu, Yu Liu, Huafeng Li, Guanqiu Qi,
Bo Xiao and Jinxing Li

Published in
Frontiers in Physics



ISSN 1664-8714
ISBN 978-2-8325-5200-1
DOI 10.3389/978-2-8325-5200-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Multi-sensor imaging and fusion: Methods, evaluations, and applications, volume II

Topic editors
Zhiqin Zhu — Chongqing University of Posts and Telecommunications, China
Yu Liu — Hefei University of Technology, China
Huafeng Li — Kunming University of Science and Technology, China
Guanqiu Qi — Buffalo State College, United States
Bo Xiao — Imperial College London, United Kingdom
Jinxing Li — Harbin Institute of Technology, Shenzhen, China

Citation
Zhu, Z., Liu, Y., Li, H., Qi, G., Xiao, B., Li, J., eds. (2024). *Multi-sensor imaging and fusion: Methods, evaluations, and applications, volume II*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5200-1

Table of contents

05 Editorial: Multi-sensor imaging and fusion: methods, evaluations, and applications, volume II
Guanqiu Qi, Zhiqin Zhu, Yu Liu, Huafeng Li, Bo Xiao and Jinxing Li

08 Dense metal corrosion depth estimation
Yanping Li, Honggang Li, Yong Guan, Xinyu Zhang and Xiaomei Zhao

22 Study on detection of intracerebral hemorrhage based on frequency difference of permittivity
Shixin Peng, Xiaoshu Wang, Gui Jin, Feng Wang, Ji Zhu, Xiaodong Zhang, Nan Liu and Rui Xu

33 Application of a digital twin for highway tunnels based on multi-sensor and information fusion
Xun Yang, Shanchuan Yu, Jun Wang, Hong Chen, Yonggang Huang, Zhongbin Luo and Lijia Fu

53 Structure similarity virtual map generation network for optical and SAR image matching
Shiwei Chen and Liye Mei

68 Hair cluster detection model based on dermoscopic images
Ya Xiong, Kun Yu, Yujie Lan, Zeyuan Lei and Dongli Fan

80 Enhanced YOLOv5s + DeepSORT method for highway vehicle speed detection and multi-sensor verification
Zhongbin Luo, Yanqiu Bi, Xun Yang, Yong Li, Shanchuan Yu, Mengjun Wu and Qing Ye

96 Auto-verbalizer filtering for prompt-based aspect category detection
Yantuan Xian, Yuan Qin and Yan Xiang

107 SCSONet: spatial-channel synergistic optimization net for skin lesion segmentation
Haoyu Chen, Zexin Li, Xinyue Huang, Zhengwei Peng, Yichen Deng, Li Tang and Li Yin

120 Cross-modality feature fusion for night pedestrian detection
Yong Feng, Enbo Luo, Hai Lu and SuWei Zhai

129 Application of mixed reality navigation technology in primary brainstem hemorrhage puncture and drainage surgery: a case series and literature review
Xiaoyong Tang, Yanglingxi Wang, Guoqiang Tang, Yi Wang, Weiming Xiong, Yang Liu, Yongbing Deng and Peng Chen

137 Brain functional magnetic resonance imaging in ICU patients who developed delirium
Ren-Jie Song, Fu-Jian Guo, Xiao-Fei Huang, Mo Li, Yi-Yun Sun, An-Yong Yu and Tian-Xi Zhang

146 **Fusion of full-field optical angiography images via gradient feature detection**
Gao Wang, Jiangwei Li, Haishu Tan and Xiaosong Li

160 **Research on electrical capacitance tomography (ECT) detection of cerebral hemorrhage based on symmetrical cancellation method**
Jing Huang, Feng Chen, Ke Wang and Sheng Chen

176 **Research on LiDAR point cloud data transformation method based on weighted altitude difference map**
Bosi Wang, Zourong Long, Xinhai Chen, Chenjun Feng, Min Zhao, Dihua Sun, Weiping Wang and Shihao Wang

186 **A survey on deep learning in medical ultrasound imaging**
Ke Song, Jing Feng and Duo Chen

Frontiers in Physics

TYPE Editorial
PUBLISHED 10 July 2024
DOI 10.3389/fphy.2024.1450685

OPEN ACCESS

EDITED AND REVIEWED BY
Cinzia Da Via,
The University of Manchester, United Kingdom

*CORRESPONDENCE
Guanqiu Qi,
qig@buffalostate.edu

RECEIVED 18 June 2024
ACCEPTED 24 June 2024
PUBLISHED 10 July 2024

CITATION
Qi G, Zhu Z, Liu Y, Li H, Xiao B and Li J (2024), Editorial: Multi-sensor imaging and fusion: methods, evaluations, and applications, volume II.
Front. Phys. 12:1450685.
doi: 10.3389/fphy.2024.1450685

Editorial: Multi-sensor imaging and fusion: methods, evaluations, and applications, volume II

Guanqiu Qi[1]*, Zhiqin Zhu[2], Yu Liu[3], Huafeng Li[4], Bo Xiao[5] and Jinxing Li[6]

[1]Computer Information Systems Department, State University of New York at Buffalo, Buffalo, NY, United States, [2]College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China, [3]Department of Biomedical Engineering, Hefei University of Technology, Hefei, China, [4]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [5]Department of Computing, Imperial College London, London, United Kingdom, [6]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

KEYWORDS

multi-sensor fusion, imaging, multi-sensor image segmentation, machine learning, performance evaluation, deep learning

Editorial on the Research Topic
Multi-sensor imaging and fusion: methods, evaluations, and applications, volume II

Introduction

Multi-sensor imaging and fusion technology plays an increasingly important role in medical imaging [1, 2], medical image segmentation [3, 4], engineering construction [5, 6], complex task object detection [7, 8] and other fields [9, 10]. Multi-sensor image fusion mainly processes images of the same object or scene captured by multiple sensors [11], which complement each other by combining multi-level and multi-spatial information, ultimately providing a consistent interpretation of the observed environment. In recent years, multi-sensor image fusion has become a highly active Research Topic, and various fusion methods have been proposed. In addition, the performance evaluation and downstream applications of multi-sensor imaging and fusion technology [12] are also receiving increasing attention. This Research Topic focuses on cutting-edge research related to multi-sensor imaging and fusion technology, including image detection and fusion methods [13], objective evaluation methods [14], and specific applications in engineering problems [15]. After a thorough peer review process, all fifteen articles submitted to this Research Topic were accepted for publication. The main research results of these works are summarized in the following three aspects.

Imaging detection, feature extraction, and fusion methods in multi-sensors

Chen et al. proposed a structure similarity virtual map generation network (SVGNet) for optical and SAR image matching. This method uses an Attention U-Net and a conditional GAN to reduce modal differences, significantly improving the matching accuracy by more than two times compared to direct image matching.

Feng et al. proposed a cross-modal fusion framework based on YOLOv5 to improve night-time pedestrian detection under low-light conditions. This dual-stream architecture processes visible and infrared images separately, using a cross-modal feature rectification module (CMFRM) to fine-tune features and reduce noise. The two-stage feature fusion module (FFM) enhances feature output through cross-attention and mixed-channel embedding, significantly improving the accuracy and robustness of night-time pedestrian detection.

Xiong et al. proposed a sparse hair cluster detection model based on improved object detection neural networks and dermoscopic images. This model utilizes a multi-level feature fusion module to extract and fuse features at different levels, and a channel-space dual attention module to enhance representation capabilities and detection accuracy. The model, tested on self-annotated data, can accurately identify and count sparse hair clusters, which outperforms existing methods in accuracy and efficiency, making it a valuable tool for early detection and treatment of hair loss.

Chen et al. proposed the spatial-channel synergistic optimization net (SCSONet), a lightweight network for skin lesion segmentation designed to run efficiently with limited computing resources. This model introduces a ConvStem module with full-dimensional attention to enhance the recognition of irregularly shaped lesion regions while reducing parameters and computational burden. The SCF block further optimizes the model by fusing spatial and channel features to reduce feature redundancy. SCSONet was validated on two public skin lesion segmentation datasets, showing high effectiveness and robustness with low computational resource requirements.

Wang et al. developed a long-depth-of-field (DOF) full-field optical angiography (FFOA) imaging system to address the limitations of capturing complete blood flow information. A novel multi-focus image fusion scheme based on gradient feature detection was proposed. This method uses non-subsampled contourlet transform (NSCT) to decompose FFOA images and applies fusion rules based on gradient feature detection. Experimental results on phantoms and animal cases showed that this method effectively expands the DOF and solves the defocus issues, providing a more comprehensive description of blood information compared with a single FFOA image.

Song et al. investigated the application of deep learning in medical ultrasound imaging with a focus on reducing computational complexity and assisting novices. They explored deep learning solutions for improving image reconstruction and clinical diagnosis.

Objective evaluation methods in multi-sensor imaging

Peng et al. developed a novel method for detecting intracerebral hemorrhage (ICH) based on the frequency-dependent variations in permittivity, eliminating the need for non-hemorrhagic baseline data. By identifying the frequency points with the maximal permittivity differences between blood and other brain tissues, the method enables absolute detection of ICH. Experimental results show that specific frequency ranges can effectively detect blood in different tissue environments, bringing promise for rapid and accurate ICH detection.

Song et al. used functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI) to detect brain alterations in intensive care unit (ICU) patients developing delirium and assess their predictive value. The study compared fifteen ICU patients with delirium to fifteen healthy controls and found significant differences in brain activity and structure. In the delirium group, the regional homogeneity (ReHo) values of the left caudate nucleus and frontal lobe were lower, the amplitude of low-frequency fluctuations (ALFF) in the hippocampus and frontal lobe was altered, and the mean diffusivity (MD), radial diffusivity (RD), fractional anisotropy (FA), and axial diffusivity (AD) in several brain regions were reduced. Early fMRI and DTI examinations are recommended to predict delirium and facilitate early intervention, potentially improving patient outcomes.

Huang et al. used electrical capacitance tomography (ECT) with a symmetrical cancellation method to detect intracerebral hemorrhage (ICH). This method places electrodes symmetrically around the head and subtracts the measured capacitances to isolate hemorrhagic events. Testing on various models shows this method can achieve absolute imaging of ICH, although mirroring artifacts and the need for precise electrode alignment pose challenges. Nonetheless, this method shows promise for pre-hospital emergency detection of ICH.

Specific applications of multi-sensor technology in engineering problems

Li et al. proposed a dense metal corrosion depth estimation method based on image segmentation and inpainting to accurately measure corrosion depth using X-ray images. This method also includes virtual data generation techniques to create training images with ground-truth annotations, thereby improving the accuracy and reliability of the corrosion depth estimates. Experimental results confirm the effectiveness of the method on both virtual and real datasets.

Yang et al. applied a digital twin to highway tunnels using a multi-modal information fusion method based on convolutional neural network (CNN)–long short-term memory (LSTM)–attention. This system solves the challenges of sensor breakdown and insufficient data support in tunnel management. By realizing closed-loop management of "accurate perception–risk assessment–decision warning–emergency management," the digital twin enhances traffic safety, reduces management costs, and improves driving comfort in highway tunnels.

Xian et al. developed an auto-verbalizer filtering method for prompt-based aspect category detection (ACD) in sentiment analysis. This approach automatically builds the verbalizer in prompt learning, enhancing the reliability of aspect categories in predictions. By leveraging the semantic extension of category labels and an indicator mechanism, their model significantly improves performance, improving by 7.5% on the zero-shot tasks and 2% on the few-shot tasks compared with the second-best models, especially excelling in handling general or miscellaneous aspect categories.

Luo et al. introduced an enhanced YOLOv5s + Deep SORT method for highway vehicle speed detection and multi-sensor verification. This approach optimizes data augmentation and incorporates the Swin Transformer module to improve local feature recognition. The model enhances vehicle detection using complete IoU (CIoU) loss for higher accuracy and Mish activation function for better convergence. Modified Deep SORT mitigates identity switching, and an image-to-coordinate transformation is used to calculate vehicle speed and average it over multiple frames. Multi-sensor verification shows the mean average precision (mAP) exceeds 90% and the speed measurement error is within 1–8 km/h, proving the model's reliability and applicability for highway scenarios.

Wang et al. developed a novel algorithm for road surface detection that combines LiDAR point clouds with 2D images to predict drivable areas for autonomous vehicle navigation. The method constructs an altitude discrepancy map from LiDAR data to exploit the height uniformity of the road surface. An innovative attention mechanism with adaptive weighting coefficients is introduced to integrate altitude disparity images with image features for semantic segmentation. Empirical evaluation using the KITTI dataset demonstrates the superior accuracy of this method in road surface detection, advancing 3Dperception technology in autonomous driving.

Tang et al. applied mixed reality navigation technology (MRNT) to brainstem hematoma puncture and drainage surgery in seven patients with primary brainstem hemorrhage (PBH). This study aims to verify the feasibility and safety of MRNT. The technology demonstrates high precision, low cost, and an immersive operating experience. The results show that the average hematoma evacuation rate was 50.39%, and the postoperative GCS scores of patients improved significantly. No intraoperative deaths or postoperative complications were reported, indicating the potential of MRNT to improve surgical outcomes in patients with PBH.

Conclusion

Overall, the Research Topic collects a wide range of relevant topics. In particular, there are research hotspots in the fields of object detection, medical image analysis and evaluation, signal monitoring and fault detection.

Special thanks to Frontier in Physics for its support and efforts in this Research Topic. We would also like to thank all authors who contributed original work to this Research Topic and all reviewers for sharing their thoughts on the submissions. We hope that this Research Topic will inspire researchers in this field and push the research on multi-sensor imaging and fusion to new frontiers.

Author contributions

GQ: Writing–original draft, Writing–review and editing. ZZ: Writing–original draft, Writing–review and editing. YL: Writing–original draft, Writing–review and editing. HL: Writing–original draft, Writing–review and editing. BX: Writing–original draft, Writing–review and editing. JL: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf Fusion* (2023) 91, 376–387.

2. Xu Y, He X, Xu G, Qi G, Yu K, Yin L, et al. A medical image segmentation method based on multi-dimensional statistical features. *Front Neurosci, Sec Brain Imaging Methods* (2022) 16. doi:10.3389/fnins.2022.1009581

3. Liu Y, Shi Y, Mu F, Cheng J, Chen X. Glioma segmentation-oriented multi-modal MR image fusion with adversarial learning. *IEEE/CAA J Autom Sin* (2022) 9(10): 1528–1531.

4. Liu Y, Mu F, Shi Y, Chen X. SF-net: a multi-task model for brain tumor segmentation in multimodal MRI via image fusion. *IEEE Signal Process Lett* (2022) 29, 1799–1803.

5. Qi G, Zhang Y, Wang K, Mazur N, Liu Y, Malaviya D. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. *Remote Sens* (2022) 14 (2), 420.

6. Zhu Z, Wang S, Gu S, Li Y, Li J, Shuai L, et al. Driver distraction detection based on lightweight networks and tiny object detection. *Math Biosci Eng* (2023) 20 (10), 18248–18266.

7. Zhu Z, Zheng R, Qi G, Li S, Li Y, Gao X. Small object detection method based on global multi-level perception and dynamic region aggregation. *IEEE Trans Circuits Syst Video Technol* (2024). doi:10.1109/TCSVT.2024.3402097

8. Li Y, Zhou Z, Qi G, Hu G, Zhu Z, Huang X. Remote sensing micro-object detection under global and local attention mechanism. *Remote Sens* (2024) 16 (4), 644.

9. Zhu Z, Wei H, Hu G, Li Y, Qi G, Mazur N. A novel fast single image dehazing algorithm based on artificial multiexposure image fusion. *IEEE Trans Instrum Meas* (2021) 70, 1–23.

10. Zheng M, Qi G, Zhu Z, Li Y, Wei H, Liu Y. Image dehazing by an artificial image fusion method based on adaptive structure decomposition. *IEEE Sens J* (2020) 20 (14), 8062–8072.

11. Zhu Z, Sun M, Qi G, Li Y, Gao X, Liu Y. Sparse dynamic volume TransUNet with multi-level edge fusion for brain tumor segmentation. *Comput Biol Med* (2024) 172, 108284.

12. Li Y, Wang Z, Yin L, Zhu Z, Qi G, Liu Y. X-net: a dual encoding–decoding method in medical image segmentation. *Vis Comput* (2023) 39, 2223–2233.

13. Qi G, Wang H, Haner M, Weng C, Chen S, Zhu Z. Convolutional neural network based detection and judgement of environmental obstacle in vehicle operation. *CAAI trans intell technol* (2019) 4 (2), 80–91.

14. Liu L, Qi Z, Cheng J, Chen X. Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: a statistic-based approach. *IEEE Trans Pattern Anal Mach Intell* (2024). doi:10.1109/TPAMI.2024.3367905

15. He X, Qi G, Zhu Z, Li Y, Cong B, Bai L. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. *Simul Model Pract Theory* (2023) 126, 102769.

TYPE Original Research
PUBLISHED 22 September 2023
DOI 10.3389/fphy.2023.1277710

Dense metal corrosion depth estimation

Yanping Li[1], Honggang Li[1], Yong Guan[2], Xinyu Zhang[2] and Xiaomei Zhao[1]*

[1]School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan, China, [2]SHI Changxu Advanced Material Innovation Center, Chinese Academy of Sciences, Shenyang, China

Check for updates

OPEN ACCESS

EDITED BY
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China

REVIEWED BY
Guanqiu Qi,
Buffalo State College, United States
Huafeng Li,
Kunming University of Science and Technology, China

*CORRESPONDENCE
Xiaomei Zhao,
✉ zhaoxiaomei20@sdjzu.edu.cn

RECEIVED 15 August 2023
ACCEPTED 11 September 2023
PUBLISHED 22 September 2023

CITATION
Li Y, Li H, Guan Y, Zhang X and Zhao X (2023), Dense metal corrosion depth estimation.
Front. Phys. 11:1277710.
doi: 10.3389/fphy.2023.1277710

Introduction: Metal corrosion detection is important for protecting lives and property. X-ray inspection systems are widely used because of their good penetrability and visual presentation capability. They can visually display both external and internal corrosion defects. However, existing X-ray-based defect detection methods cannot present and estimate the dense corrosion depths. To solve this problem, we propose a dense metal corrosion depth estimation method based on image segmentation and inpainting.

Methods: The proposed method employs an image segmentation module to segment metal corrosion defects and an image inpainting module to remove these segmented defects. It then calculates the pixel-level dense corrosion depths using the X-ray images before and after inpainting. Moreover, to address the difficulty of acquiring training images with ground-truth dense corrosion depth annotations, we propose a virtual data generation method for creating virtual corroded metal X-ray images and their corresponding ground-truth annotations.

Results: Experiments on both virtual and real datasets show that the proposed method successfully achieves accurate dense metal corrosion depth estimation.

Discussion: In conclusion, the proposed virtual data generation method can provide effective and sufficient training samples, and the proposed dense metal corrosion depth estimation framework can produce accurate dense corrosion depths.

KEYWORDS

corrosion depth estimation, image segmentation, image inpainting, virtual training data generation, x-ray image

1 Introduction

Metal objects are common and important in daily life. However, contact with air and water often cause unavoidable corrosion during the service life of metal components. Corrosion significantly reduces the strength of metal materials, shortening their service life and even posing serious safety hazards. Therefore, timely and accurate metal corrosion detection can effectively protect lives and property.

At present, many defect detection methods have been proposed, using RGB [1] or RGB-D images [2, 3], eddy currents [4], and ultrasound [5]. However, these methods either cannot detect internal corrosion defects or cannot visually display them. In contrast, X-ray inspection systems have the visual presentation capability to display both external and internal structures. Therefore, X-ray inspection systems are often used to detect metal defects including corrosion. Existing automatic defect detection methods using X-ray images fall

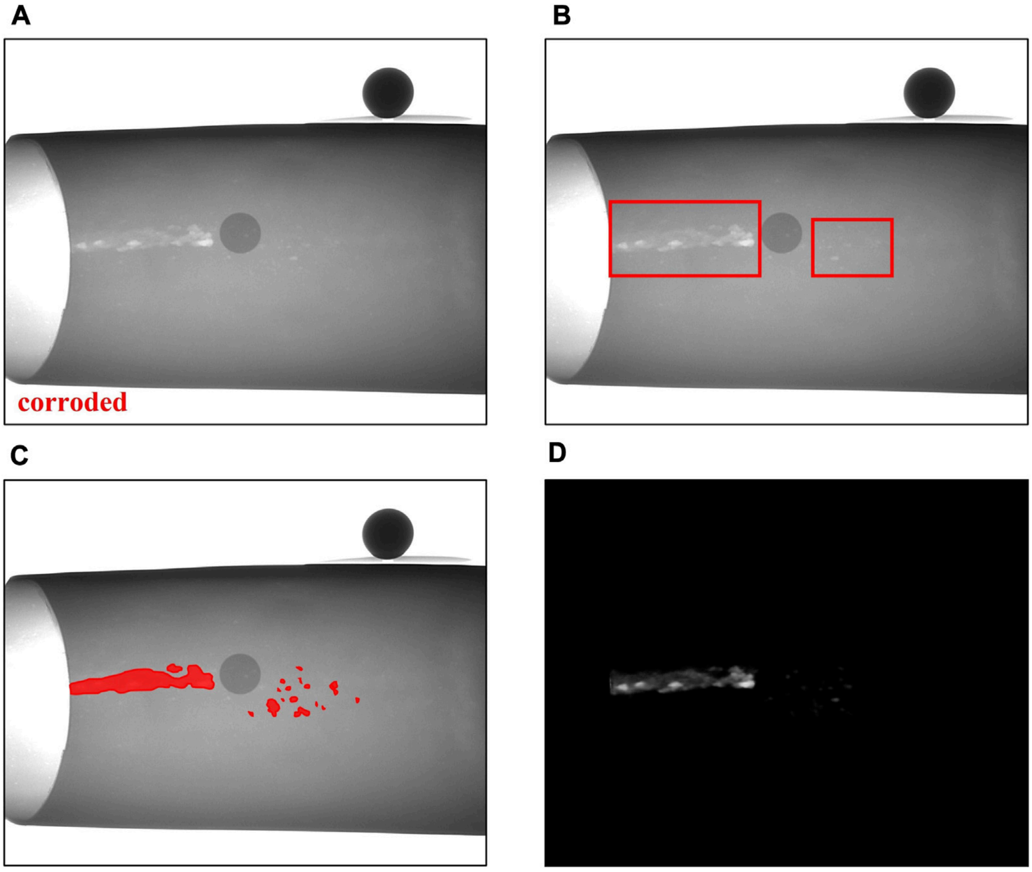


FIGURE 1
Comparison of results of different kinds of defect detection methods **(A)** classification-based method **(B)** target detection-based method **(C)** segmentation-based method; and **(D)** the proposed dense metal corrosion depth estimation method.

into three categories: classification-based, target detection-based, and segmentation-based methods.

1) Classification-based methods generally use and improve classic classification networks [6–9] such as Inception [10] and VGG [11]. For example, Zhang et al. [9] trained Inception and MobileNet [12] by transfer learning and combined these two networks through a multi-module ensemble framework to classify weld defects. Hu et al. [6] proposed an object-level attention mechanism and used this mechanism to train a VGG16-based type classification module and a defect classification module to classify casting defects. Jiang et al. [8] improved VGG16 by employing attention-guided data augmentation to train the casting defect classification network with effective data augmentation. Tang et al. [7] improved VGG16 by employing a spatial attention mechanism and bilinear pooling to classify casting defects. As shown in Figure 1A, classification methods only output an image-level classification result to determine whether there is a defect in the image.
2) Target detection-based methods generally use and improve popular object detection networks such as Faster RCNN [13]. For example, Gong et al. [14] improved domain adaptive Faster RCNN (DA Faster) [13] by adding a feature pyramid network (FPN) [15], small anchor strategies, ROI Align, and other strategies to detect defects in spacecraft composite structures. Liu et al. [16] improved Faster RCNN by employing a residual network combined with FPN and an efficient convolutional attention module to detect weld defects. Cheng et al. [17] improved DS-Cascade RCNN [18] by adding a spatial attention mechanism, deformable convolution and pruning algorithms to detect wheel hub defects. As shown in Figure 1B, these target detection methods can roughly locate the position of defects using bounding-boxes.
3) Segmentation-based methods generally use segmentation networks with encoder–decoder structures, such as U-Net [19]. Du et al. [20] improved U-Net to segment defects in casting parts by changing its backbone to ResNet 101 [21], adding a contrast-limited adaptive histogram equalization module, a gated multi-layer fusion module, and a weighted intersection over union (IOU) loss function. Yang et al. [22] improved U-Net by adding a multi-scale feature fusion block and a bidirectional convolutional Long Short-Term Memory block to segment welding defects. Du et al. [23] built an interactive X-ray network (IXNet) with a click attention module based on U-Net to perform interactive segmentation of casting defects. As shown in Figure 1C, the segmentation results of these methods contain detailed defect location, area, and shape information.

Of all these methods, segmentation-based approaches provide the most detailed defect information. Despite this, even these

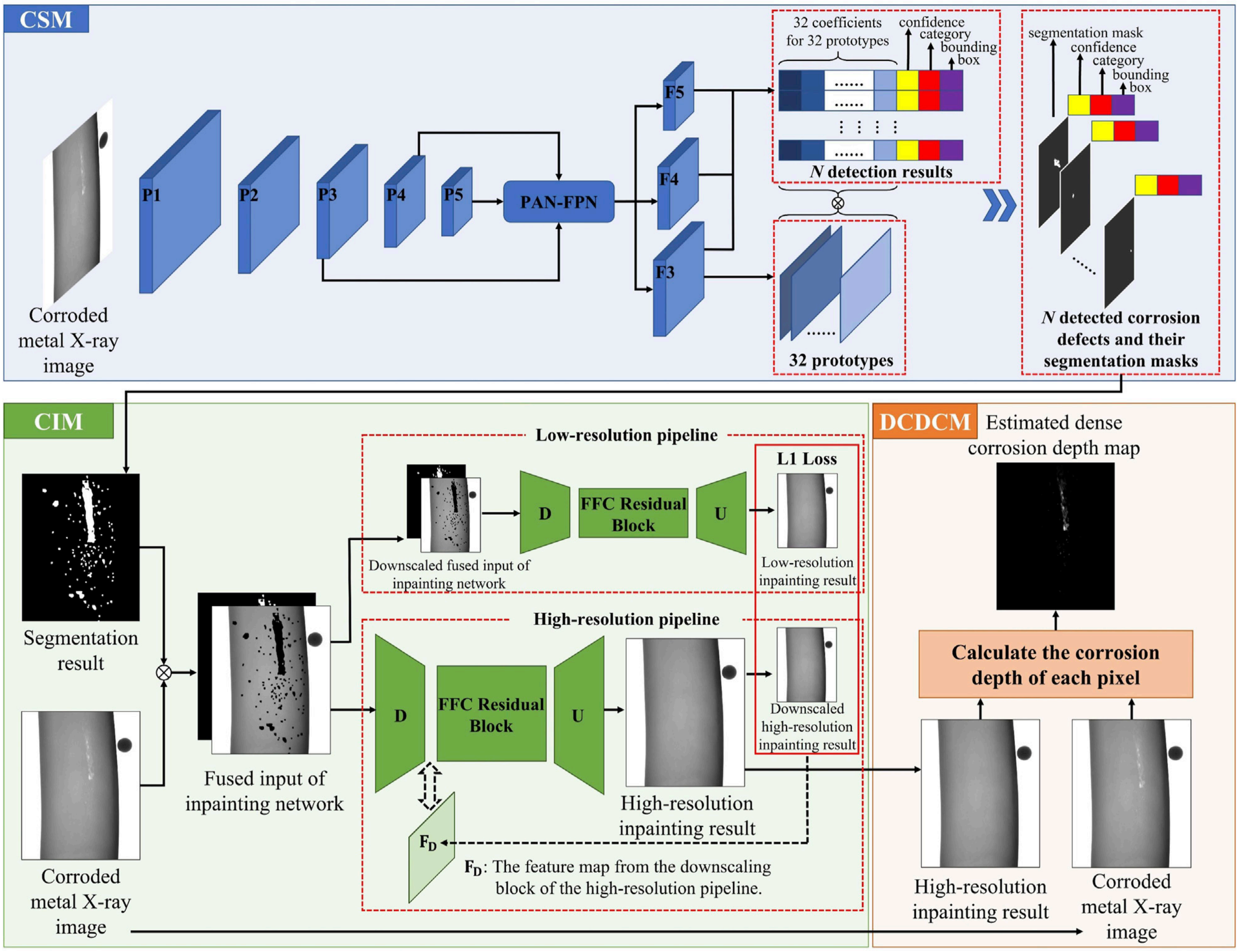

CSM
Corroded metal X-ray image
P1
P2
P3
P4
P5
PAN-FPN
F5
F4
F3
32 coefficients for 32 prototypes
confidence
category
bounding box
N detection results
32 prototypes
segmentation mask
confidence
category
bounding box
N detected corrosion defects and their segmentation masks
CIM
Segmentation result
Corroded metal X-ray image
Fused input of inpainting network
Low-resolution pipeline
Downscaled fused input of inpainting network
D
FFC Residual Block
U
L1 Loss
Low-resolution inpainting result
High-resolution pipeline
D
FFC Residual Block
U
High-resolution inpainting result
Downscaled high-resolution inpainting result
F_D
F_D: The feature map from the downscaling block of the high-resolution pipeline.
DCDCM
Estimated dense corrosion depth map
Calculate the corrosion depth of each pixel
High-resolution inpainting result
Corroded metal X-ray image

FIGURE 2
The flow chart of our proposed dense metal corrosion depth estimation framework. CSM denotes the corrosion segmentation module. CIM denotes the corrosion inpainting module. DCDCM denotes the dense corrosion depth calculation module.

methods cannot estimate the depth of defects, which is a crucial parameter for metal corrosion analysis. Currently, software developed by NOVO DR Ltd. Is able to estimate the depth of defects through the Double Wall Technique (DWT) [24]. The major drawback of DWT is its strongly reliance on manual operation, making it unusable for automatic depth estimation.

To address these limitations, we propose a new defect detection method that detects corrosion defects based on dense metal corrosion depth estimation. The proposed method is capable of automatically estimating the corrosion depth maps that contain dense corrosion depth information. An example of estimated corrosion depth map from our method is shown in Figure 1D, with the value of each pixel denoting its corresponding corrosion depth. The estimated corrosion depth map not only contains dense corrosion depth information but also includes detailed information regarding the location, area, and shape of corrosion defects.

The proposed method is composed of three modules for corrosion segmentation, corrosion inpainting, and dense corrosion depth calculation. The corrosion segmentation module is based on the state-of-the-art real-time instance segmentation method YOLOv8 [25]. The corrosion inpainting module is based on the state-of-the-art image inpainting method LAMA [26]. The corrosion depth calculation module is based on the Beer–Lambert law [27]. Both the corrosion segmentation and the corrosion inpainting modules are based on deep learning neural networks, which require a large number of training images with ground-truth annotations. However, annotating corrosion defects in X-ray images not only requires significant manpower and time but also extensive expertise. This implies that only adequately trained researchers possess the ability to annotate corrosion defects in X-ray images. As a result, it is very hard to annotate sufficient X-ray images for training. To address this issue, we propose a novel virtual data generation method. This method can generate virtual corroded metal X-ray images and their corresponding ground-truth annotations automatically without any manual intervention.

The main contributions of this paper are as follows.

1) We propose a novel dense metal corrosion depth estimation framework to address the problem that previous technologies cannot automatically estimate dense corrosion depths. This proposed framework uses a corrosion segmentation module

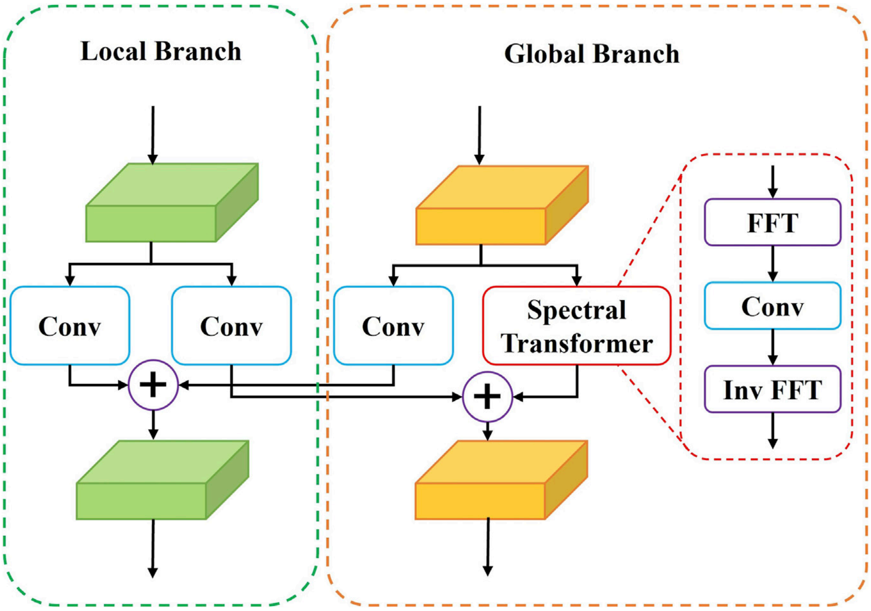


FIGURE 3
The architecture of fast Fourier convolution (FFC).

(CSM) to segment corrosion defects and a corrosion inpainting module (CIM) to remove these segmented corrosion defects. Then, a dense corrosion depth calculation module (DCDCM) is employed to calculate the pixel-level dense corrosion depths using the X-ray images before and after inpainting.

2) We propose a novel virtual data generation method to address the issue that it is difficult to manually annotate dense corrosion depths in X-ray images. This proposed method contains a virtual corrosion cell generation module (VCCGM) to generate virtual corrosion cells, and a virtual corrosion image generation module (VCIGM) to generate virtual corroded metal X-ray images and their corresponding ground-truth dense corrosion depth annotations. With the help of this method, sufficient virtual images and their ground-truth annotations are generated for training and testing.
3) We perform sufficient experiments on both virtual and real datasets to prove the effectiveness of the proposed virtual data generation method and dense metal corrosion depth estimation framework. The experimental results show that the proposed framework trained by the generated virtual dataset successfully produces accurate dense metal corrosion depths.

2 Dense metal corrosion depth estimation

2.1 Overview

The process flow of our dense metal corrosion depth estimation framework is shown in Figure 2. The framework is composed of three modules: the corrosion segmentation module (CSM), the corrosion inpainting module (CIM), and the dense corrosion depth calculation module (DCDCM). An incoming X-ray image with corrosion defects is first given to CSM. CSM outputs its corresponding corrosion segmentation result. CIM then removes the corrosion defects according to the original X-ray image and its corresponding corrosion segmentation result. Finally, DCDCM calculates the corrosion depth of each pixel according to the X-ray images before and after inpainting. These modules are described in detail in the following sections.

2.2 Corrosion segmentation module

In the field of computer vision, YOLO plays an important role. It stands out from a large number of methods for its remarkable balance of speed and accuracy [28]. The first version of YOLO was proposed in 2015 [29]. Through the efforts of many researchers, the eighth version of YOLO, YOLOv8, was proposed in early 2023 [25]. YOLOv8 achieves state-of-the-art performance in real-time object detection and instance segmentation. Therefore, we use YOLOv8 in our corrosion segmentation module (CSM).

The simplified network architecture of YOLOv8 is shown within the CSM in Figure 2. As shown, five convolutional blocks are first employed to extract high-level features. After passing through each convolutional block, the height and width of feature map are reduced. In Figure 2, these feature maps produced by the different convolutional blocks are denoted as P_1, P_2, P_3, P_4, and P_5. Then, a neck block called PANFPN is employed to combine image features from P_3, P_4 and P_5, enhancing the spatial and semantic information across different scales. PANFPN outputs three collections of features, each at different scale, denoted as F_3, F_4, and F_5. The heights and widths of F_3, F_4, and F_5 match the heights and widths of P_3, P_4, and P_5, respectively. Finally, the category, bounding box, and segmentation mask of each object are predicted using F_3, F_4, and F_5.

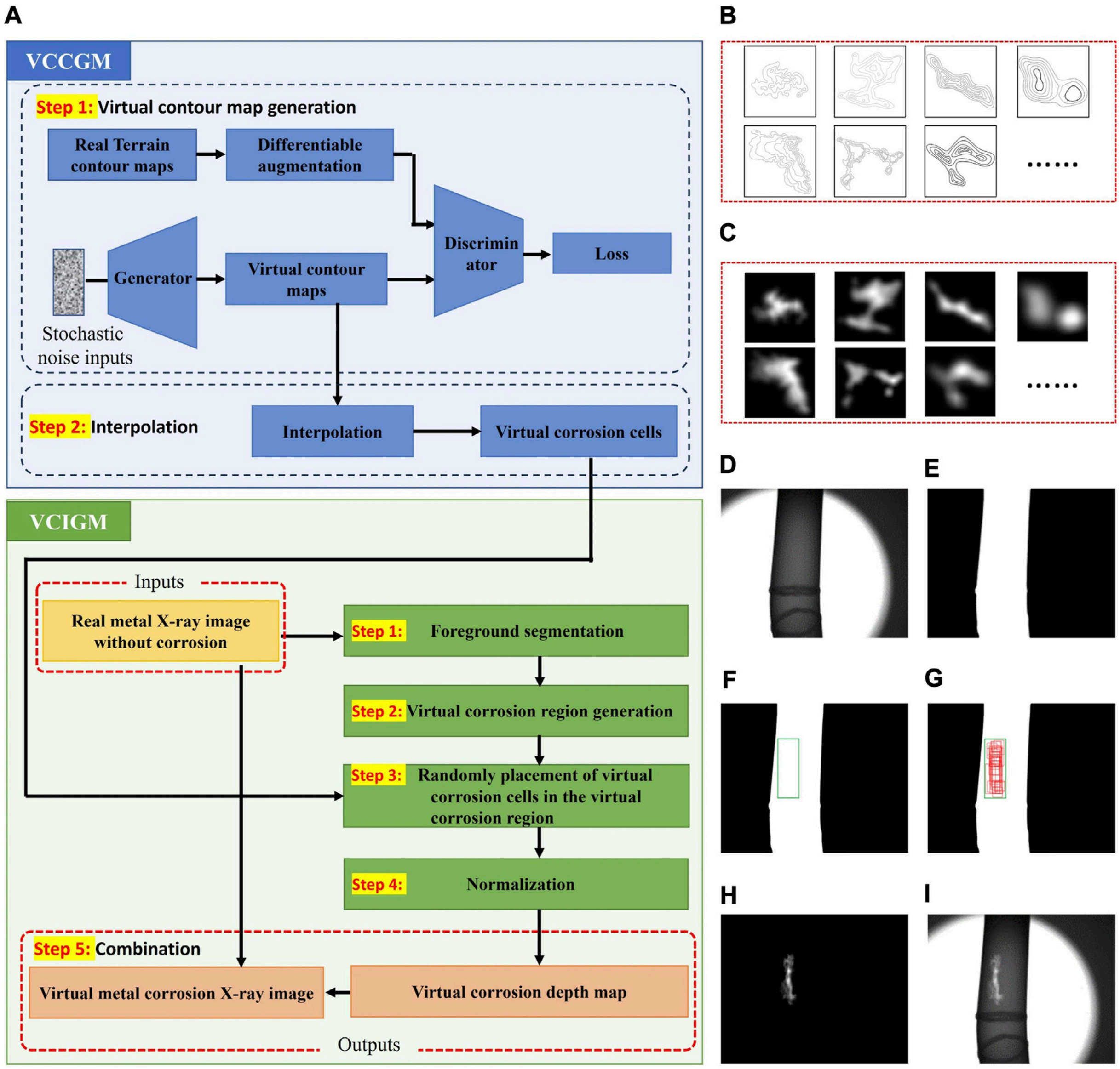

A
VCCGM
Step 1: Virtual contour map generation
Real Terrain contour maps
Differentiable augmentation
Generator
Virtual contour maps
Discriminator
Loss
Stochastic noise inputs
Step 2: Interpolation
Interpolation
Virtual corrosion cells
VCIGM
Inputs
Real metal X-ray image without corrosion
Step 1: Foreground segmentation
Step 2: Virtual corrosion region generation
Step 3: Randomly placement of virtual corrosion cells in the virtual corrosion region
Step 4: Normalization
Step 5: Combination
Virtual metal corrosion X-ray image
Virtual corrosion depth map
Outputs
B
C
D
E
F
G
H
I
......

FIGURE 4
The components of the virtual data generation method and examples **(A)** the flowchart of the VCCGM (virtual corrosion cell generation module) and the VCIGM (virtual corrosion image generation module) **(B)** examples of virtual contour maps **(C)** examples of virtual corrosion cells **(D)** a real metal X-ray image without corrosion **(E)** the foreground segmentation result **(F)** the generated virtual corrosion region **(G)** the randomly selected regions that used to place virtual corrosion cells **(H)** the generated virtual corrosion depth map; and **(I)** the generated virtual corroded metal X-ray image.

TABLE 1 Evaluation scores of the framework with different instance segmentation models.

Frameworks with different instance segmentation models	mAP_{50}^{box}	mAP_{50}^{mask}	mIoU (%)	Speed (ms)	MAE ($\times10^{-2}$) ↓	MSE ($\times10^{-2}$) ↓
Framework with YOLOv5	73.6%	61.1%	62.6	38.3	1.32	2.26
Framework with YOLOv7	73.4%	60.3%	62.3	37.5	1.33	2.37
Framework with YOLOv8	**75.0%**	**71.3%**	**69.4**	38.5	**1.23**	**1.92**

Scores marked in bold indicate the best results on the corresponding metric.

We design the CSM module to segment corrosion defects from X-ray images of corroded metal materials. As shown in Figure 2, the outputs of the neural network consist of two parts: N detection results and 32 segmentation prototypes [30]. Let us use Pre_N to denote the prediction results, where $Pre_N = \{pre_n | n = 1, 2, 3, \ldots\ldots, N\}$, N is the number of detected corrosion defects, and pre_n is the n^{th} detection

TABLE 2 Evaluation scores of the proposed framework with different inpainting models.

Frameworks with different inpainting models	MAE ($\times 10^{-2}$) ↓	MSE ($\times 10^{-2}$) ↓
Framework with AOT	7.05	189.51
Framework with PUT	3.99	26.74
Framework with LAMA	**1.23**	**1.92**

Scores marked in bold indicate the best results on the corresponding metric.

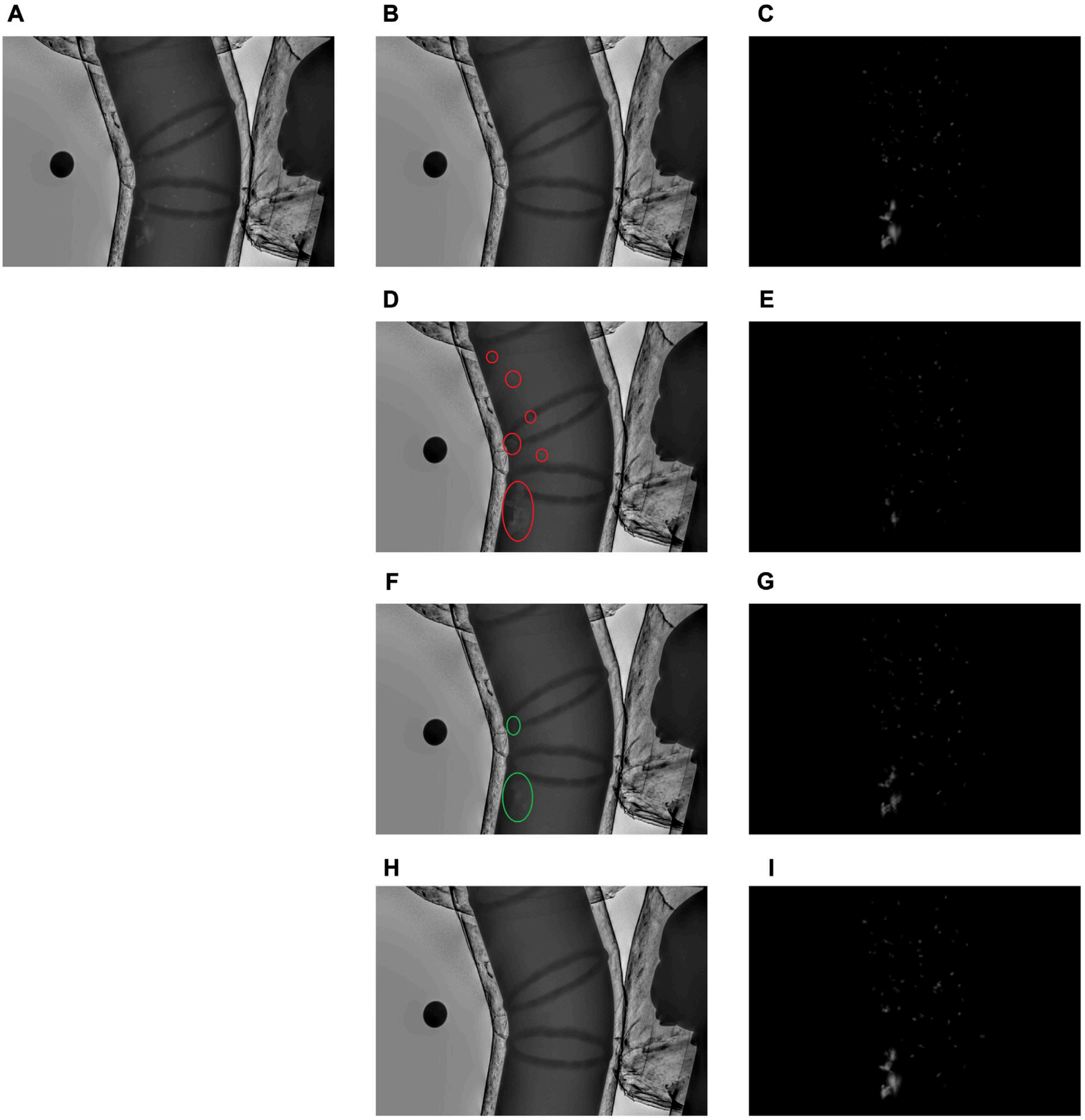


FIGURE 5
The inpainting and depth map estimation results of different frameworks **(A)** virtual corroded metal X-ray image **(B)** the ground-truth inpainting result **(C)** the ground-truth depth map **(D)** the inpainting result of AOT **(E)** the estimated depth map of the framework with AOT **(F)** the inpainting result of PUT **(G)** the estimated depth map of the framework with PUT **(H)** the inpainting result of LAMA; and **(I)** the estimated depth map of the framework with LAMA.

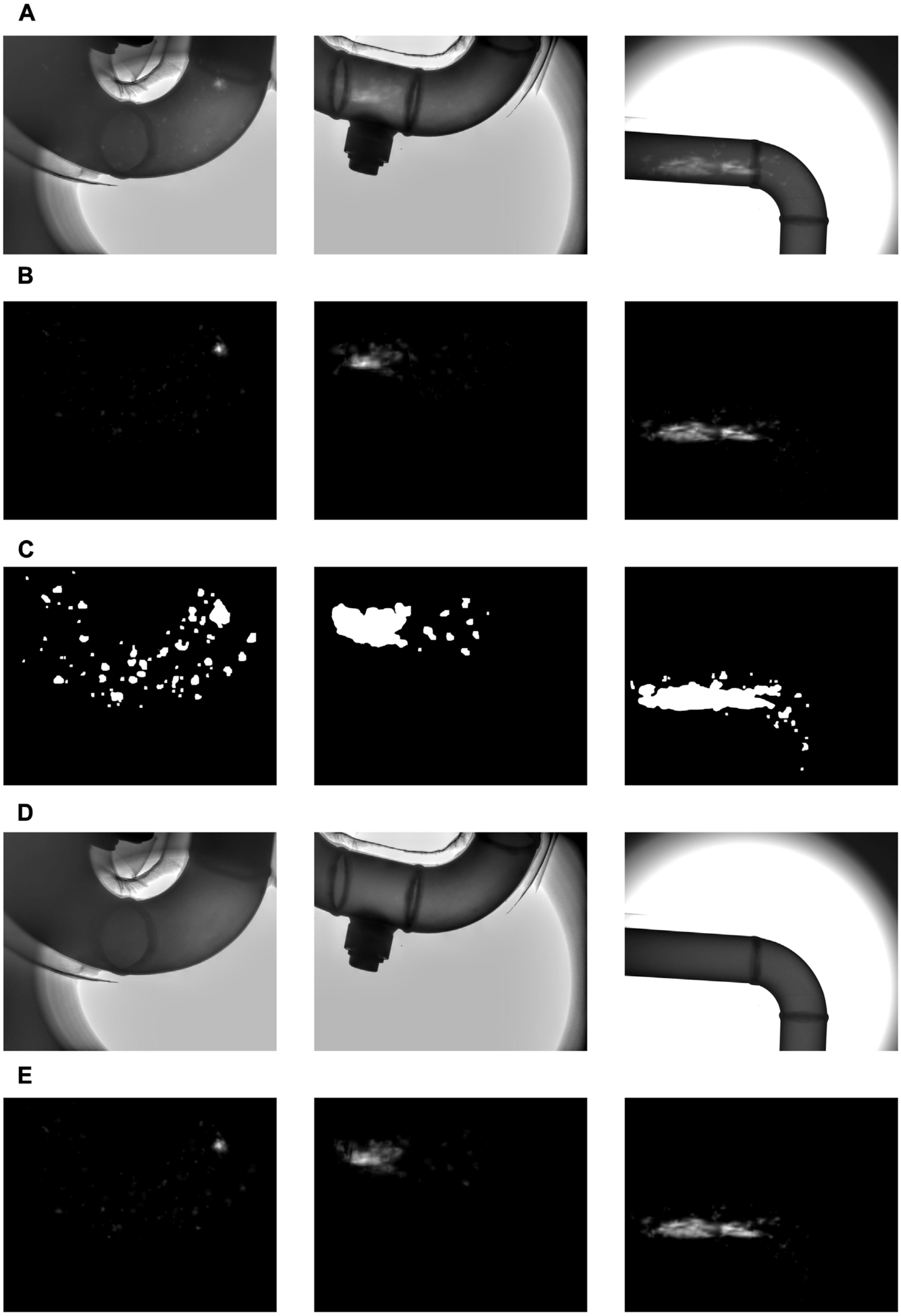


FIGURE 6
Examples of dense metal corrosion depth estimation on virtual images **(A)** the virtual corroded metal X-ray image **(B)** the ground-truth depth maps **(C)** the corrosion defect segmentation results of CSM (using YOLOv8) **(D)** the corrosion defect inpainting results of CIM (using LAMA); and **(E)** the estimated corrosion depth map.

result. $pre_n = \left\{Coef_n^{32}, Conf_n^1, Clas_n^1, Box_n^4\right\}$. $Coef_n^{32}$ denotes the segmentation prototype coefficients in the n^{th} detection result, and $Coef_n^{32} = \left\{coef_n^l | l = 1, 2, 3, \ldots\ldots, 32\right\}$, where $coef_n^l$ denotes the l^{th} segmentation prototype coefficient. $Conf_n^1$ denotes the confidence of the n^{th} detection result. The length of $Conf_n^1$ is 1. $Clas_n^1$ denotes the classification result of the n^{th} detection result. The length of $Clas_n^1$ is 1.

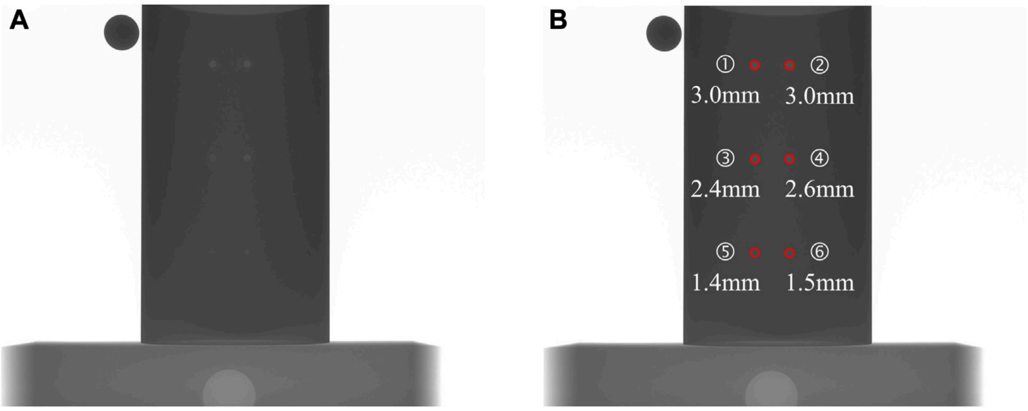


FIGURE 7
X-ray images of the test metal pipe **(A)** raw image of the metal pipe with six holes of known depths; and **(B)** annotated image with the position and depth of each hole.

TABLE 3 The depth estimation results of the proposed framework and DWT using a real X-ray image of a metal pipe with six holes of known depths.

Index of holes	Ground-truth depth (mm)	DWT		Ours	
		Predicted depth (mm)	Absolute error (mm)	Predicted depth (mm)	Absolute error (mm)
①	3.00	3.08	0.08	3.04	0.04
②	3.00	3.16	0.16	3.18	0.18
③	2.40	2.24	0.16	2.27	0.13
④	2.60	2.49	0.11	2.52	0.08
⑤	1.40	1.12	0.28	1.11	0.29
⑥	1.50	1.22	0.28	1.21	0.29

Box_n^4 denotes the bounding box of the n^{th} detection result. The length of Box_n^4 is 4, and it contains the horizontal and vertical coordinates of the upper-left corner of the bounding box, as well as the width and height of the bounding box. The 32 segmentation prototypes are denoted as $Pro^{32} = \{pro^l | l = 1, 2, 3, \ldots\ldots, 32\}$, where pro^l denotes the l^{th} segmentation prototype. The segmentation mask of the n^{th} detection result is calculated based on $Coef_n^{32}$, Box_n^4, and Pro^{32} through the following three steps:

(1) $Coef_n^{32} = \left\{coef_n^l | l = 1, 2, 3, \ldots\ldots, 32\right\}$ are used as the combination weights to linearly combine the 32 segmentation prototypes $Pro^{32} = \{pro^l | l = 1, 2, 3, \ldots\ldots, 32\}$ and obtain the combination result com_n, $com_n = \sum_{l=1}^{32} coef_n^l \times pro^l$.
(2) com_n is processed using a sigmoid nonlinearity operation and a binarization operation to obtain the primary segmentation mask $pm_n = Binary(Sigmoid(com_n))$, where $Sigmoid()$ denotes the sigmoid nonlinearity operation and $Binary()$ denotes the binarization operation.
(3) The primary segmentation mask pm_n is cropped by the bounding box of the n^{th} detection result Box_n^4, and the final segmentation mask of the n^{th} detection result $m_n = Crop(pm_n)$ is obtained. The cropping operation $Crop()$ assigns zero to pixels outside of Box_n^4.

A set of corrosion segmentation results are shown in Figure 2. Each detected corrosion defect contains its bounding box coordinates, classification value, confidence value, and segmentation mask.

During training of the CSM, binarized virtual corrosion depth maps and the bounding boxes of disconnected corrosion areas are used as the ground truth of the instance segmentation results. Further details on the generation of virtual corroded X-ray images and their corresponding ground-truth depth maps are presented in Section 3.

2.3 Corrosion inpainting module

In the proposed dense corrosion depth estimation framework, we use a corrosion inpainting module to remove corrosion defects. This module employs the state-of-the-art image inpainting method LAMA [26]. LAMA builds its inpainting network using fast Fourier convolutions (FFCs) to obtain an image-wide receptive field and improve inpainting performance.

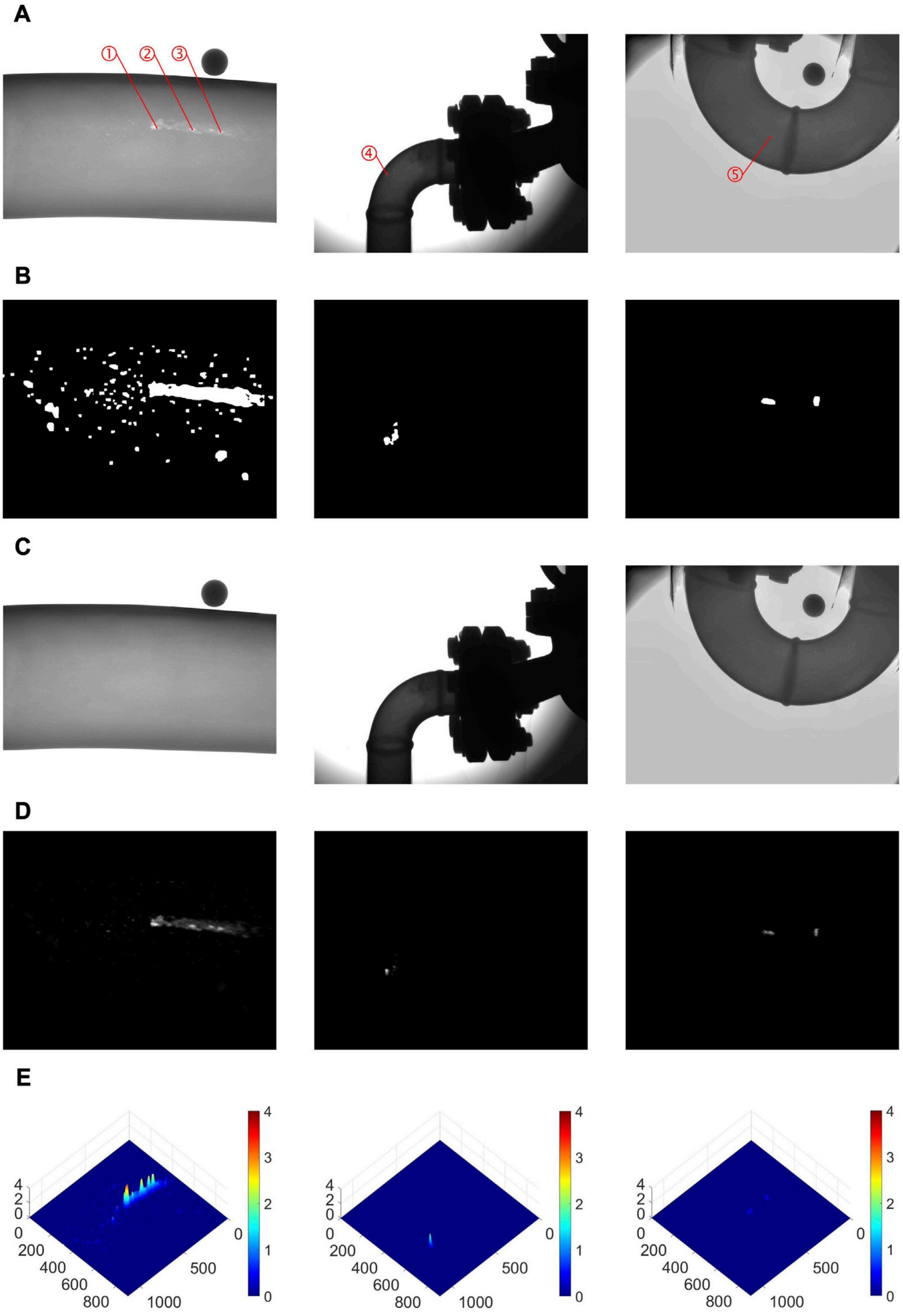


FIGURE 8
Examples of dense metal corrosion depth estimation with real images **(A)** the real corroded metal X-ray image **(B)** the corrosion defect segmentation results from CSM (using YOLOv8) **(C)** the corrosion defect inpainting results from CIM (using LAMA) **(D)** the estimated corrosion depth maps (to make the corrosion defects more significant, each corrosion depth map has been divided by its maximum depth); and **(E)** the estimated corrosion depth map shown in 3D.

The architecture of FFC is shown in Figure 3. FFC contains two parallel branches: a local branch and a global branch [31]. The local branch uses conventional convolutions to extract local features. The global branch uses a spectral transformer to extract global features. The spectral transformer first transforms image features into a spectral domain by fast Fourier transform (FFT), then conducts

TABLE 4 Corrosion depth estimation results for our framework and DWT using real corroded metal X-ray images.

Index of points	DWT (mm)	Ours (mm)
①	3.15	3.06
②	2.66	2.55
③	2.23	2.16
④	1.77	1.79
⑤	0.68	0.79

an efficient global update in the spectral domain, and finally converts features back to the spatial domain via inverse fast Fourier transform (Inv FFT). A point-wise update in the spectral domain globally affects all spatial features [31]. Therefore, the spectral transformer can extract global features. The above local features and global features are then combined to fuse multi-scale features. A large effective receptive field plays a crucial role in the inpainting task [26]. However, conventional convolutions cannot provide a large effective receptive field, especially in the early layers of the network. In contrast, FFC can provide an image-level receptive field in very early layers of the network [26]. Therefore, FFC effectively improves the inpainting performance.

The simplified network architecture of LAMA is shown within the CIM in Figure 2. In our project, the resolution of the processed images is large, so we use an architecture containing low-resolution and high-resolution pipelines. These two pipelines use the same network (i.e., having the same architecture and weights) to process inputs in different resolutions. As shown in Figure 2, the inpainting network contains three blocks: a downscaling block (labeled D in Figure 2), an FFC residual block, and an upscaling block (labeled U in Figure 2). The downscaling block contains 3 FFCs with strides set to 2. The FFC residual block contains 18 sub-residual blocks built on FFCs with strides set to 1. The upscaling block uses 3 transpose convolutions with strides set to 2. The inpainting results produced by the inpainting network have the same size as the inputs as a result.

In the inpainting architecture, the two pipelines play different roles. As shown in Figure 2, the low-resolution pipeline uses the downscaled inputs for inpainting. Smaller inputs are beneficial to generate inpainting results with better global structures [32]. However, many image details can be lost during the down-sampling operation. In contrast, the high-resolution pipeline uses the original inputs for inpainting. No image details are lost in its inputting step. However, larger inputs cause incoherent structures [32]. To maintain image details while generating inpainting results with better global structures, the inpaining results of the low-resolution pipeline are used to supervise the global structures of the inpainting results of the high-resolution pipeline. The supervision process is operated by minimizing the L1 loss between the downscaled high-resolution inpainting results and the low-resolution inpainting results. Note that the L1 loss is minimized by updating the feature map from the downscaling block of the high-resolution pipeline F_D (as shown in Figure 2), rather than the parameters in the neural network. Using the above method, F_D can learn the global structures of the low-resolution inpainting results. F_D passes these good global structures to the final inpainting results of the high-resolution pipeline through forward propagation. Therefore, the final inpainting results of the high-resolution pipeline can maintain image details and have good global structures.

The binary mask used for inpainting is provided by the CSM, as shown in Figure 2. It covers all detected corrosion defects. The original X-ray image, as shown in Figure 2, is masked using this binary mask. This masked X-ray image is then stacked with the binary mask to generate a fused input. The inpainting network finally outputs the inpainting result with the same scale as the original X-ray image, as shown in Figure 2. A comparison of the images in Figure 2 before and after inpainting shows that the corrosion defects have been removed.

When training CIM, the actual X-ray images without corrosion that used to combine virtual corrosion depth maps are used as the ground truth of image inpainting results. Further details on the generation of virtual corroded X-ray images and their corresponding ground-truth depth maps are presented in Section 3.

2.4 Dense corrosion depth calculation module

In X-ray images, the gray value of each pixel is exponentially related to the corresponding thickness of the transilluminated material as given by:

$$g_k = g_k^o e^{-\mu t_k} \tag{1}$$

where g_k represents the gray value of the k^{th} pixel in X-ray image $I = \{g_k | k = 1, 2, 3, \ldots, K\}$; K denotes the total number of pixels in this image; g_k^o is a parameter related to the intensity of incident X-ray; μ represents the attenuation coefficient, which can be roughly considered as a constant when the material category and the radiation source are the same; and t_k represents the thickness of the corresponding transilluminated material. If corrosion occurs and the corrosion depth is Δt_k, the gray value will change to:

$$g_k^c = g_k^o e^{-\mu(t_k - \Delta t_k)} \tag{2}$$

Eq. 2 can be rewritten as:

$$g_k^c = g_k^o e^{-\mu t_k} \cdot e^{-\mu \Delta t_k} \tag{3}$$

As $g_k^o e^{-\mu t_k} = g_k$, we obtain the equation:

$$g_k^c = g_k \cdot e^{\mu \Delta t_k} \tag{4}$$

According to Eq. 4, the corrosion depth can be calculated as:

$$\Delta t_k = \frac{1}{\mu} \ln \left(\frac{g_k^c}{g_k} \right) \tag{5}$$

Therefore, when the values of μ, g_k^c, and g_k are known, the corrosion depth of the k^{th} pixel Δt_k can be calculated. The value of μ can be calibrated by a step wedge of the same material in advance. g_k^c is the gray value of the k^{th} pixel in the corroded metal X-ray image $I^c = \{g_k^c | k = 1, 2, 3, \ldots, K\}$, and I^c is the X-ray image to be processed. g_k is the gray value of the k^{th} pixel in the X-ray image without corrosion defects I. In practice, when we obtain the X-ray image to be processed I^c, it is difficult to obtain its corresponding I. In this paper, we use the inpainting result

$\widetilde{I^c} = \left\{\widetilde{g_k^c} | k = 1, 2, 3, \ldots, K\right\}$, instead of the real I. The estimated corrosion depth of the k^{th} pixel $\Delta\widetilde{t_k}$ can then be calculated as:

$$\Delta\widetilde{t_k} = \frac{1}{\mu}\ln\left(\frac{g_k^c}{\widetilde{g_k^c}}\right) \tag{6}$$

From Eq. 6, we obtain a pixel-level dense corrosion depth map $\Delta\widetilde{T} = \{\Delta\widetilde{t_k} | k = 1, 2, 3, \ldots, K\}$, as shown in Figure 2.

3 Virtual data generation

As described above, the proposed dense corrosion depth estimation framework contains two deep learning-based modules: CSM and CIM, both of which need a large number of annotated images for training. However, it is quite difficult to annotate the corrosion defects in real X-ray images. We propose the virtual data generation method in this section to solve this problem. This method automatically generates virtual corroded metal X-ray images and their corresponding virtual corrosion depth maps for the purpose of acquiring sufficient and various annotated training X-ray images automatically.

A flowchart of the proposed virtual data generation method is shown in Figure 4A. A set of images in the key steps of our proposed method are shown in Figures 4(B)–(I). This method consists of two modules: the virtual corrosion cell generation module (VCCGM) and the virtual corrosion image generation module (VCIGM). VCCGM and VCIGM cooperate to generate the virtual corrosion image and the corresponding corrosion depth map. Specifically, VCCGM provides virtual corrosion cells for VCIGM; VCIGM randomly combines these virtual corrosion cells to generate virtual corrosion depth maps and combines the virtual corrosion depth maps with real metal X-ray images without corrosion to generate virtual corroded metal X-ray images.

In the following subsections, we first introduce the working principle of the virtual data generation method in detail, and then we introduce VCCGM and VCIGM in detail.

3.1 Principle of virtual data generation

As shown in Eq. 4, when the gray value without corrosion g_k, the corrosion depth Δt_k, and the attenuation coefficient μ are known, we can obtain the gray value after corrosion g_k^c. g_k comes from I, the X-ray image without corrosion, with $I = \{g_k | k = 1, 2, 3, \ldots, K\}$ and K denoting the total number of pixels in the image. We can obtain I by taking X-ray images of metal materials without corrosion. Based on I, if we wish to obtain a virtual g_k^c, denoted as $\widehat{g_k^c}$, we need to generate a virtual Δt_k, denoted as $\Delta\widehat{t_k}$, and a virtual μ, denoted as $\hat{\mu}$:

$$\widehat{g_k^c} = g_k \cdot e^{\hat{\mu}\Delta\widehat{t_k}} \tag{7}$$

In this paper, we treat $\hat{\mu}$ as a constant, generated by experience. Thus, the challenge of generating $\widehat{g_k^c}$ is how to generate $\Delta\widehat{t_k}$. The values of $\Delta\widehat{t_k}$ differ for each pixel, but the values of $\Delta\widehat{t_k}$ are not independent within the image. These values have a reasonable global structure. Therefore, instead of generating pixel-level $\Delta\widehat{t_k}$ values one by one, we generate an image-level virtual dense corrosion depth map $\Delta\hat{T} = \{\Delta\widehat{t_k} | k = 1, 2, 3, \ldots, K\}$. The relationship among I, $\hat{\mu}$, $\Delta\hat{T}$, and $\widehat{I^c} = \left\{\widehat{g_k^c} | k = 1, 2, 3, \ldots, K\right\}$ can be formulated as:

$$\widehat{I^c} = I \cdot e^{\hat{\mu}\Delta\hat{T}} \tag{8}$$

Therefore, the main mission of the proposed virtual data generation method is to generate a reasonable virtual dense corrosion depth map $\Delta\hat{T}$. $\Delta\hat{T}$ is then combined with the existing real X-ray image without corrosion I to generate the virtual corroded metal X-ray image $\widehat{I^c}$.

In the field of image processing, a generative adversarial network (GAN) is commonly used for generating virtual images. However, a GAN needs a large number of real data samples for training, and it is difficult to acquire real dense corrosion depth maps. Therefore, it is difficult to train a GAN that can generate virtual dense corrosion depth maps.

Through the observation of many corroded metal X-ray images, we find that the brightness fluctuations in the corrosion areas of X-ray images are similar to the topographic fluctuations. Thus, one solution to the above problem is to borrow the concept of contour maps from geography and use terrain contour maps downloaded from the internet as real data samples to train a GAN that can generate virtual contour maps. Then, virtual depth maps can be obtained by interpolating these virtual contour maps. However, the above solution has two problems: 1) it is difficult to download sufficient complex terrain contour maps to simulate complex corrosions, and 2) it is difficult to interpolate complex contour maps.

To solve the above two problems, we only use a GAN to generate virtual corrosion cells by VCCGM, and then we randomly combine different virtual corrosion cells to generate virtual corrosion depth maps by VCIGM. Although it is difficult to download sufficient complex terrain contour maps, it is much easier to download simple terrain contour maps with one or two peaks. We use these simple terrain contour maps downloaded from the internet as real data samples to train a GAN that can generate simple virtual contour maps. Virtual corrosion cells can be generated by interpolating these simple virtual contour maps. By randomly combining different virtual corrosion cells, we can generate a large number of various virtual corrosion depth maps.

3.2 Virtual corrosion cell generation module

The virtual corrosion cell generation module is designed to generate a series of virtual corrosion cells as shown in the examples in Figure 4C. Virtual corrosion cells are sub-depth maps with simple structures and fixed scale.

In order to generate sufficient virtual corrosion cells with a variety of structures, we create the virtual corrosion cell generation module (VCCGM) using a generative adversarial network (GAN). GAN is a common method used for data augmentation. A simplified GAN structure is shown in the VCCGM of Figure 4A. GAN has two main blocks: a generator block and a discriminator block. During training, the two blocks play against each other and finally generate virtual data samples which are indistinguishable from real ones.

Even though GAN is able to generate a large number of high-quality virtual data samples, it also needs a large number of real data samples for training. However, it is very difficult to obtain a sufficient number of real corrosion cells. To solve this problem, we borrow the concept of contour maps from geography and use some terrain contour maps downloaded from the Internet as real data samples.

The processing steps of VCCGM are as follows.

Step 1. Virtual contour map generation. The GAN, which is trained by terrain contour maps, generates virtual contour maps, with examples shown in Figure 4B.

Step 2. Interpolation. The generated virtual contour maps are interpolated to generate virtual corrosion cells as shown in Figure 4C.

3.3 Virtual corrosion image generation module

The virtual corrosion image generation module generates virtual corrosion depth maps by combining the virtual corrosion cells provided by the VCCGM and generates virtual corroded metal X-ray images by combining the generated virtual corrosion depth maps with real X-ray images without corrosion. The flow chart of this module has been shown in the VCIGM in Figure 4A. To ensure that the generated virtual corrosion defects locate at the foreground areas, the X-ray images without corrosion that used to combine virtual corrosion depth maps also participate in generating virtual corrosion depth maps. The steps of how to use virtual corrosion cells to generate virtual corrosion depth maps and how to generate virtual corroded metal X-ray images are as follows.

Step 1. Foreground segmentation. An actual X-ray image without corrosion, as shown in Figures 4D, is sent into the foreground segmentation step. The foreground segmentation part, built using YOLOv8, produces the segmentation result shown in Figure 4E;

Step 2. Virtual corrosion region generation. This step randomly generates a bounding box in the segmented foreground area. The green bounding box in the white foreground area shown in Figure 4F is an example. Virtual corrosion will be put in this bounding box;

Step 3. Random placement of virtual corrosion cells in the virtual corrosion region. A cluster of sub-boxes are randomly generated in the virtual corrosion region. These sub-boxes have been marked in red in Figure 4G. They have different sizes and different aspect ratios. Each sub-box selects a virtual corrosion cell generated by VCCGM and resizes the selected virtual corrosion cell to fill itself. If overlap occurs, the overlapping parts are added together. After this step, a preliminary virtual corrosion depth map is obtained;

Step 4. Normalization. This step normalizes the preliminary virtual corrosion depth map into a reasonable value range. The upper bound of corrosion depth equals the thickness of inspected metal material. The lower bound of corrosion depth is 0. The max value of depth map d_{max} is randomly selected between the upper and lower bounds. Then, the value range of preliminary virtual depth map is linearly transformed to $[0, d_{max}]$ to obtain the final virtual corrosion depth map shown in Figure 4H;

Step 5. Combination. This step combines the generated virtual corrosion depth map shown in Figure 4H and the actual X-ray image without corrosion shown in Figure 4D according to Eq. 8. The result is a virtual corroded metal X-ray image, as shown in Figure 4I.

4 Experiments

In this paper, we have presented our framework for estimating the dense metal corrosion depth using X-ray images. In view of the previously described difficulties in obtaining actual corroded metal X-ray images with ground-truth annotations, we have also presented a method for generating virtual corrosion images for the purposes of training our method. In our experiments, we used 16,199, 4,200, and 2,170 virtual images for training, validation, and testing, respectively. To verify that the model trained on virtual datasets is also suitable for real datasets, we also tested our proposed model on several real cases. All our experiments were implemented using PyTorch with two NVidia RTX 3090 GPUs and one Intel Xeon Gold 5222 CPU.

4.1 Experiments on virtual dataset

As described in Section 2, our framework has three modules: a corrosion segmentation module (CSM), a corrosion inpainting module (CIM), and a dense corrosion depth calculation module (DCDCM). CSM and CIM use the YOLOv8 real-time instance segmentation model and the LAMA inpainting model, respectively. To verify their effectiveness, we also performed experiments with other models. We employed mean absolute error (MAE) and mean square error (MSE) to evaluate the corrosion depth estimation performance. The formulas of MAE and MSE are:

$$MAE = \frac{1}{K}\sum_{k=1}^{K}\left|\left(\Delta t_k^{gt} - \Delta t_k^{p}\right)\right| \tag{9}$$

$$MSE = \frac{1}{K}\sum_{k=1}^{K}\left(\Delta t_k^{gt} - \Delta t_k^{p}\right)^2 \tag{10}$$

where K denotes the total number of pixels in this image; Δt_k^{gt} represents the corrosion depth value of the k^{th} pixel in the ground-truth depth map; and Δt_k^{p} represents the corrosion depth value of the k^{th} pixel in the predicted depth map. The evaluation scores of the proposed framework with different instance segmentation models and different inpainting models are shown in Table 1 and Table 2.

To compare different instance segmentation models in more aspects, we also show mAP_{50}^{box}, mAP_{50}^{mask} [25], mIoU [33], and processing speed in Table 1. As shown in this table, we tested three instance segmentation models (YOLOv5 [34], YOLOv7 [35], and YOLOv8 [25]) with the proposed framework. The use of YOLOv8 yielded the best performance, largely owing to its higher segmentation accuracy. The processing speeds of these three instance segmentation methods are comparable.

As shown in Table 2, we tested three inpainting models (AOT [32], PUT [36], and LAMA) on the proposed framework. LAMA provided the best performance, with a large performance gap compared to the others, because the inpainting performance of LAMA is significantly higher than that of the other two methods. To qualitatively compare the inpainting performance of AOT, PUT, and LAMA, we show a group of their inpainting results in Figure 5.

As shown in Figure 5, the corrosion regions are still readily visible (indicating reduced inpainting performance) in the inpainting result of AOT as indicated by the red circles. PUT was better, but in the regions marked with green circles, the differences between corroded and normal areas are still visible. In the inpainting results of LAMA, it is quite difficult to distinguish corrosion regions from normal regions. As LAMA provides the best inpainting results, the corrosion depths calculated from its inpainting results are more accurate. The predicted corrosion depth maps of the proposed frameworks with AOT, PUT, and LAMA are also shown in Figure 5. The predicted corrosion depth map when using LAMA is closest to the ground-truth depth map.

Figure 6 shows more examples of corrosion depth estimation with virtual cases. CSM accurately segmented most corrosion defects; CIM successfully removed the segmented corrosion defects; and DCDCM estimated accurate and reasonable depth maps that are fairly close to the ground-truth depth maps.

4.2 Experiments on real dataset

It is extremely difficult to quantitatively evaluate the dense metal corrosion depth estimation performance on real corroded metal X-ray images because it is hard to obtain their ground-truth corrosion depth maps. In this section, we used a metal pipe with six holes of known depths to quantitatively evaluate the depth estimation accuracy of our proposed framework, and collected several real corroded metal X-ray images to qualitatively evaluate the dense metal corrosion depth estimation performance of our proposed framework.

4.2.1 Quantitative experiment

As noted, we used a metal pipe with six holes of known depths. The wall thickness of this pipe was 3 mm. The raw and annotated X-ray images are shown in Figure 7. The depth estimation results of our framework and DWT are shown in Table 3. DWT is the defect depth estimation method used in NOVO DR systems [24].

As shown in Table 3, the depth estimation absolute errors of our framework are comparable with DWT, indicating similar accuracy. DWT, however, requires human–computer interaction, while our framework is fully automatic. Therefore, our framework is more convenient to use.

4.2.2 Qualitative experiment

In this experiment, we collected some real X-ray images of corroded metal pipes. Because we could not obtain their ground-truth corrosion depth maps, we could not quantitatively evaluate the dense corrosion depth estimation performance using MAE and MSE. However, we can still qualitatively analyze the performance of our proposed framework by checking whether the estimated corrosion depth maps are reasonable. Three group of examples are shown in Figure 8.

As shown in Figure 8, CSM successfully segmented the corrosion defects; CIM successfully removed these corrosion defects, obtaining accurate inpainting results; and DCDCM successfully estimated the dense corrosion depth maps according to the original corroded metal X-ray images and their corresponding inpainting results. In order to present the estimated corrosion depth maps more vividly, they are shown in 3D in the last row of Figure 8. As shown in Figure 8A, the three cases had different degrees of corrosion: in the first case, the corrosion defects were large, dense, and deep; in the second case, the corrosion defects were much smaller; in the third case, the corrosion defects were very shallow. The estimated corrosion depth maps shown in Figure 8E are consistent with these observations.

Even though we could not obtain the ground truth of corrosion depths, we still compared the depth estimation results of our framework with DWT at five different points, labeled as ①, ②, ③, ④, and ⑤ in Figure 8. The depth estimation results of these five points are shown in Table 4. DWT is widely used in the NOVO DR systems [24]. Although it is not perfect (as shown in Table 3, the depth estimation results of DWT are not exactly equal to the ground-truth values), widespread experience shows that DWT is a reliable defect depth estimation method. As shown in Table 4, the depth estimation results of our framework are close to the depth estimation results of DWT, demonstrating that the depth values in our estimated dense corrosion depth maps are reasonable.

5 Conclusion

In this paper, we propose a novel dense metal corrosion depth estimation framework for X-ray images. It consists of three modules: a corrosion segmentation module (CSM), a corrosion inpainting module (CIM), and a dense corrosion depth calculation module (DCDCM). CSM segments corrosion defects from the X-ray images. CIM removes these segmented corrosion defects. DCDCM calculates the corrosion depth maps, which contain dense corrosion depth information, according to the original X-ray images and the inpainting results of CIM. To solve the problem of lacking training dataset with ground-truth of annotations, we propose a virtual data generation method to generate virtual corroded metal X-ray images and their corresponding ground-truth corrosion depth annotations. The virtual data generation method consists of two modules: a virtual corrosion cell generation module (VCCGM) and a virtual corrosion image generation module (VCIGM). VCCGM generates virtual corrosion cells using a generative adversarial network. VCIGM generates virtual corrosion depth maps by combining the virtual corrosion cells and produces virtual corroded metal X-ray images by combining the generated virtual corrosion depth maps with actual X-ray images without corrosion. We use these generated images to train both CSM and CIM. Experimental results show that the proposed dense metal corrosion depth estimation framework trained using the generated virtual dataset could successfully estimate accurate and dense metal corrosion depth automatically.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

YL: Formal Analysis, Methodology, Software, Writing–original draft. HL: Formal Analysis, Methodology, Writing–original draft, Software. YG: Formal Analysis, Writing–review and editing. XnZ: Formal Analysis, Writing–review and editing. XaZ: Formal Analysis, Writing–review and editing, Funding acquisition, Methodology, Project administration, Software, Writing–original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Natural Science Foundation of Shandong Province (ZR2021QF094) and the Youth Innovation Team Technology Project of Higher School in Shandong Province (2022KJ204).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Wu J, Zhou W, Qiu W, Yu L. Depth repeated-enhancement rgb network for rail surface defect inspection. *IEEE Signal Process. Lett* (2022) 29:2053–7. doi:10.1109/LSP.2022.3211199

2. Wang J, Song K, Zhang D, Niu M, Yan Y. Collaborative learning attention network based on rgb image and depth image for surface defect inspection of no-service rail. *IEEE/ASME Trans Mechatronics* (2022) 27(6):4874–84. doi:10.1109/TMECH.2022.3167412

3. Zhou W, Hong J. Feet: Lightweight feature hierarchical exploration network for real-time rail surface defect inspection in rgb-d images. *IEEE Trans Instrumentation Meas* (2023) 72:1–8. doi:10.1109/TIM.2023.3237830

4. Farag H E, Toyserkani E, Khamesee MB. Non-destructive testing using eddy current sensors for defect detection in additively manufactured titanium and stainless-steel parts. *Sensors* (2022) 22(14):5440. doi:10.3390/s22145440

5. Yu Y, Safari A, Niu X, Drinkwater B, Horoshenkov KV. Acoustic and ultrasonic techniques for defect detection and condition monitoring in water and sewerage pipes: A review. *Appl Acoust* (2021) 183:108282. doi:10.1016/j.apacoust.2021.108282

6. Hu C, Wang Y. An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images. *IEEE Trans Ind Electron* (2020) 67(12):10922–30. doi:10.1109/TIE.2019.2962437

7. Tang Z, Tian E, Wang Y, Wang L, Yang T. Nondestructive defect detection in castings by using spatial attention bilinear convolutional neural network. *IEEE Trans Ind Inform* (2021) 17(1):82–9. doi:10.1109/TII.2020.2985159

8. Jiang L, Wang Y, Tang Z, Miao Y, Chen S. Casting defect detection in x-ray images using convolutional neural networks and attention-guided data augmentation. *Measurement* (2021) 170:108736. doi:10.1016/j.measurement.2020.108736

9. Zhang H, Chen Z, Zhang C, Xi J, Le X. Weld defect detection based on deep learning method. In: Proceedings of the IEEE International Conference on Automation Science and Engineering; August 2019; Vancouver, BC, Canada (2019). p. 1574–9. doi:10.1109/COASE.2019.8842998

10. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proc AAAI Conf Artif Intelligence* (2017) 31(1):4278–84. doi:10.1609/aaai.v31i1.11231

11. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition (2014). Avaialble at: https://arxiv.org/abs/1409.1556. doi:10.48550/arXiv.1409.1556

12. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017). Avaialble at: https://arxiv.org/abs/1704.04861. doi:10.48550/arXiv.1704.04861

13. Ren S, He K, Girshick R, SunFaster JR-CNN. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Machine Intelligence* (2017) 39(6):1137–49. doi:10.1109/TPAMI.2016.2577031

14. Gong Y, Luo J, Shao H, Li Z. A transfer learning object detection model for defects detection in x-ray images of spacecraft composite structures. *Compos Structures* (2022) 284:115136. doi:10.1016/j.compstruct.2021.115136

15. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; July 2017; Honolulu, HI, USA (2017). p. 936–44. doi:10.1109/CVPR.2017.106

16. Liu W, Shan S, Chen H, Wang R, Sun J, Zhou Z. X-ray weld defect detection based on AF-RCNN. *Welding In The World* (2022) 66(6):1165–77. doi:10.1007/s40194-022-01281-w

17. Cheng S, Lu J, Yang M, Zhang S, Xu Y, Zhang D, et al. Wheel hub defect detection based on the DS-Cascade RCNN. *Measurement* (2023) 206:112208. doi:10.1016/j.measurement.2022.112208

18. Cai Z, Vasconcelos N, Cascade R-CNN. Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 2018; Salt Lake City, UT, USA (2018). p. 6154–62. doi:10.1109/CVPR.2018.00644

19. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference; Proceedings, Part III 18. 2015; October, 2015; Munich, Germany. Springer (2015). p. 234–41. doi:10.1007/978-3-319-24574-4_28

20. Du W, Shen H, Fu J. Automatic defect segmentation in x-ray images based on deep learning. *IEEE Trans Ind Electron* (2021) 68(12):12912–20. doi:10.1109/TIE.2020.3047060

21. 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; June 2016; Las Vegas, NV, USA (2016). p. 770–8. doi:10.1109/CVPR.2016.90

22. Yang L, Song S, Fan J, Huo B, Li E, Liu Y. An automatic deep segmentation network for pixel-level welding defect detection. *IEEE Trans Instrumentation Meas* (2022) 71:1–10. doi:10.1109/TIM.2021.3127645

23. Du W, Shen H, Zhang G, Yao X, Fu J. Interactive defect segmentation in x-ray images based on deep learning. *Expert Syst Appl* (2022) 198:116692. doi:10.1016/j.eswa.2022.116692

24. NOVO DR Ltd. *Novo portable digital radiography system user manual* (2021).

25. Jocher G, Chaurasia A, Qiu J. Yolo by ultralytics (2023). Avaialble at: https://github.com/ultralytics/ultralytics.

26. Suvorov R, Logacheva E, Mashikhin A, Remizova A, Ashukha A, Silvestrov A, et al. Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; January 2022; Waikoloa, HI, USA (2022). p. 3172–82. doi:10.1109/WACV51458.2022.00323

27. Hsieh J. *Computed tomography: Principles, design, artifacts, and recent advances* (2003).

28. Jiang P, Ergu D, Liu F, Cai Y, Ma B. A review of yolo algorithm developments. *Proced Comput Sci* (2022) 199:1066–73. doi:10.1016/j.procs.2022.01.135

29. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 2016; Las Vegas, NV, USA (2016). p. 779–88. doi:10.1109/CVPR.2016.91

30. Bolya D, Zhou C, Xiao F, Lee YJ. Yolact++ better real-time instance segmentation. *IEEE Trans Pattern Anal Machine Intelligence* (2022) 44(2):1108–21. doi:10.1109/TPAMI.2020.3014297

31. Chi L, Jiang B, Mu Y. Fast fourier convolution. In: *Proceedings of advances in neural information processing systems* (2020). Cambridge: MIT Press, p. 4479–88.

32. Zeng Y, Fu J, Chao H, Guo B. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Trans Visualization Comput Graphics* (2023) 29(7):3266–80. doi:10.1109/TVCG.2022.3156949

33. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation (2017). Avaialble at: https://arxiv.org/abs/1704.06857. doi:10.48550/arXiv.1704.06857

34. Jocher G. Yolov5 by ultralytics (2020). Avaialble at: https://github.com/ultralytics/yolov5.

35. Wang C-Y, Bochkovskiy A, Liao H-YM. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 2022; New Orleans, LA, USA (2022). p. 7464–75. doi:10.48550/arXiv.2207.02696

36. Liu Q, Tan Z, Chen D, Chu Q, Dai X, Chen Y, et al. Reduce information loss in transformers for pluralistic image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 2022; New Orleans, LA, USA (2022). p. 11337–47. doi:10.1109/CVPR52688.2022.01106

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 19 December 2023
DOI 10.3389/fphy.2023.1333185

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Yiyao An,
Chongqing University of Posts and Telecommunications, China
Shuai Zhou,
Kunming University of Science and Technology, China
Dahong Gao,
henna finance university, China

*CORRESPONDENCE
Nan Liu,
natasha0902@sina.com
Rui Xu,
xurui203389@hospital.cqmu.edu.cn

[†]These authors have contributed equally to this work

RECEIVED 04 November 2023
ACCEPTED 23 November 2023
PUBLISHED 19 December 2023

CITATION
Peng S, Wang X, Jin G, Wang F, Zhu J, Zhang X, Liu N and Xu R (2023), Study on detection of intracerebral hemorrhage based on frequency difference of permittivity.
Front. Phys. 11:1333185.
doi: 10.3389/fphy.2023.1333185

Study on detection of intracerebral hemorrhage based on frequency difference of permittivity

Shixin Peng[1†], Xiaoshu Wang[1†], Gui Jin[2], Feng Wang[2], Ji Zhu[1], Xiaodong Zhang[1], Nan Liu[3*] and Rui Xu[1*]

[1]Department of Neurosurgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [2]College of Biomedical Engineering, Third Military Medical University, Army Medical University, Chongqing, China, [3]Department of Radiology, Chongqing Red Cross Hospital, Jiangbei District People's Hospital, Chongqing, China

Introduction: Current detection of intracerebral hemorrhage (ICH), whether employing Electrical Capacitance Tomography (ECT) or other electrical imaging techniques, rely on time-difference measurements. The time-difference methods necessitate baseline measurements from the patient in a non-hemorrhagic state, which is impractical to obtain, rendering rapid detection of ICH unfeasible.

Methods: This study introduces a novel approach that capitalizes on the distinct dispersion characteristics of the permittivity in brain tissue and the spectral variance of the permittivity between blood and other brain components. Specifically, the frequency-dependent variations in the permittivity are employed to achieve absolute detection of ICH, thereby eliminating the need for non-hemorrhagic baseline data. The methodology entails identification of two frequency points that the frequency-dependent variation in the permittivity at these two frequency points manifest the maximal difference between blood and other brain tissues. Subsequently, this permittivity differential at the two identified frequency points is utilized for hemorrhage detection. Experimental measurements were conducted using an impedance analyzer and a parallel plate capacitor to capture the capacitance in four single-component substances—distilled water, sheep blood, isolated pig fat, and isolated pig brain—as well as three mixed blood compounds—distilled water enveloping sheep blood, pig fat encapsulating sheep blood, and pig brain surrounding sheep blood—across a frequency range of 10 kHz to 20 MHz.

Results: The results show that in different frequency bands, it is indeed possible to distinguish single-component substances from mixed substances by the frequency difference of capacitance variation. Comparative analysis reveals that the 1 MHz to 5 MHz frequency range is most effective for detecting blood in distilled water. For blood detection in pig fat, a 10 kHz to 1 MHz frequency range is identified as optimal, while a 10 kHz to 0.5 MHz frequency range is advantageous for blood detection in pig brain tissue.

Discussion: The findings confirm that absolute detection of ICH is achievable through frequency-dependent variations in the permittivity. However, this necessitates the identification of the frequency band manifesting the largest difference of frequency-dependent variation between single-component and mixed substances. The study acknowledges limitations primarily due to the use of anticoagulant-altered sheep blood, which exhibits permittivity divergent from those of natural blood. Additionally, the *in vitro* pig fat and pig brain samples, having been subjected to freeze-thaw cycles, also demonstrate permittivity unrepresentative of *in vivo* tissue.

KEYWORDS

intracerebral hemorrhage, ECT, permittivity, frequency-differential, capacitance

Introduction

Intracerebral hemorrhage (ICH) is an acute cerebrovascular disorder precipitated by the rupture of cerebral vessels, leading to blood accumulation within the brain parenchyma [1]. Characterized by rapid onset and elevated mortality, ICH poses a significant public health concern. Among cerebrovascular events, it is second only to ischemic stroke in incidence, with a rate of 12–15 cases per 100,000 person-years [2]. In Western countries, ICH comprises approximately 15% of all stroke cases, whereas in China, this figure varies between 18% and 47% [3]. The ailment exhibits a substantial 30-day mortality rate of 35%–52%. Moreover, a mere 20% of affected individuals regain self-care capabilities within a 6-month period, thereby exerting a considerable socioeconomic burden. ICH can be classified into primary and secondary types; primary ICH accounts for 80%–85% of all instances and is predominantly associated with hypertension, thereby termed as hypertensive intracerebral hemorrhage. Presently, China's hypertensive population approximates 245 million, establishing hypertension as the principal risk factor for ICH [4]. Timely identification and intervention are pivotal for enhancing therapeutic success and postoperative outcomes. For cases with minimal bleeding and relative clinical stability, intervention within 6 hours post-onset is recommended. In contrast, for cases with extensive bleeding and critical status, emergency intervention within 1 hour is imperative. Currently, computed tomography (CT) scanning is considered the diagnostic gold standard for ICH. However, substantial delays occur in the period between patient transport to the hospital, CT examination, and the receipt of diagnostic results. These delays often lead to a missed therapeutic window for effective ICH treatment. Furthermore, the bulky nature of existing diagnostic equipment precludes its use for pre-hospital triage and bedside monitoring. Consequently, there is an urgent need for a portable, cost-effective, non-invasive, and rapid detection technology for ICH.

The detection of ICH using the electrical characteristics of biological tissues is a new type of measurement technology, especially represented by Electrical Impedance Tomography (EIT) and Magnetic Induction Tomography (MIT) [5, 6]. Studies on the permittivity of brain tissues have indicated a significantly higher permittivity for blood compared to other cerebral constituents. Specifically, at a frequency of 1 MHz, the permittivity values for blood, gray matter, and cerebrospinal fluid are 3,000, 990, and 108, respectively [7]. Although permittivity across all brain tissues declines as frequency increases, blood consistently exhibits elevated permittivity levels. Therefore, theoretically, monitoring alterations in cerebral permittivity offers a more effective approach to ICH detection. Electrical Capacitance Tomography (ECT) serves as a technique to map the permittivity distribution within an object, reliant on multi-electrode capacitance measurements. Originally developed for applications in oil industry multiphase flow measurements and fluidized bed evaluations within the pharmaceutical sector [8, 9], ECT has demonstrated promise in ICH detection. In prior experiments, a parallel-plate capacitor was employed to directly measure capacitance changes within the brain due to hemorrhage. Animal experiments revealed a positive correlation between injected blood volume and induced cerebral capacitance alterations [10]. Subsequently, a 16-channel ECT system was developed and successfully utilized for *in vitro* imaging of hemorrhagic events within porcine cerebral tissue [11].

The outcomes of these experiments substantiate the feasibility of utilizing capacitance changes in brain tissue as a marker for the detection of ICH. While the latter experiment successfully employed ECT for *in vitro* ICH imaging, the methodology was reliant on time-difference imaging. Specifically, this involved subtracting reference data obtained prior to the hemorrhagic event from post-bleed measurements, an approach commonly employed in contemporary electrical imaging modalities [12, 13]. However, this time-difference detection strategy necessitates baseline measurements from the patient prior to the onset of ICH, which is operationally challenging. As such, this method is unsuitable for rapid ICH detection and is limited to monitoring temporal variations in ICH. To achieve rapid ICH detection, it is imperative to obtain absolute distribution data, akin to the information provided by CT and MRI. Traditional electrical imaging faces several limitations in this context. First, the subtle differences in permittivity between ICH-affected tissue and normal brain tissue are difficult to discern, particularly given the diminutive volume of ICH relative to overall brain tissue. Consequently, the weak electrical or magnetic signals emitted from hemorrhagic regions are often masked by signals from other normal brain tissues, thereby complicating the separation of weak ICH signals from dominant background signals [14]. Second, the sensitivity of traditional electrical imaging techniques for biological tissue detection is markedly inferior to that of CT and MRI. These challenges collectively hinder the application of traditional electrical imaging for the acquisition of absolute ICH distribution within the brain, relegating it to the role of monitoring dynamic ICH changes over time.

Due to the dispersion characteristics inherent in biological tissues, they exhibit varying electrical properties across different frequencies [15]. Gabrel et al. assessed the electrical conductivity and permittivity distribution of diverse brain tissues within the frequency range of 10 Hz–20 GHz [16]. The findings revealed that although the permittivity of all brain tissues decreases with increasing frequency, the rate of this decrease varies among tissues and frequency bands. Importantly, the permittivity of blood consistently surpassed that of other brain tissues across all

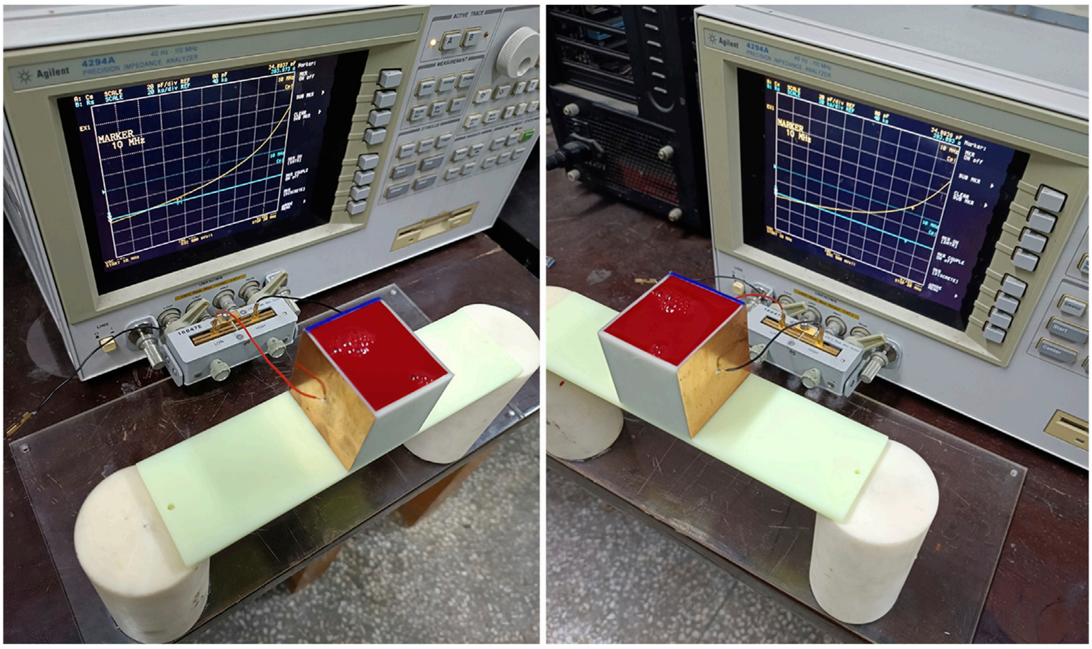


FIGURE 1
Capacitance measurement system based on a parallel plate capacitor.

frequencies, displaying a larger range of change in specific frequency bands. Consequently, these changes of permittivity with frequency can be harnessed for the detection of ICH. To address the limitations of time-difference methods, the current study employed a parallel plate capacitor coupled with an impedance analyzer. This configuration facilitated the measurement of capacitance distribution in both blood and normal brain tissue, within a frequency span of 10 kHz–20 MHz. The measured capacitance is proportional to the permittivity of the substance under test. The objective was to identify a frequency band where the difference of permittivity frequency change between ICH-affected and normal brain tissues is maximized. Subsequently, capacitance changes in this identified frequency band were analyzed for both ICH and non-ICH models. The aim was to evaluate the feasibility of leveraging these differences for ICH detection. To validate the applicability of this approach, the permittivity frequency spectrum of various substances—namely distilled water, sheep blood, pig fat, and pig brain—was measured. Utilizing the frequency-dependent differences in permittivity, we conducted assessments on three distinct models: blood suspended in distilled water, blood encased in pig fat, and blood embedded in pig brain. The results of these experiments are expected to corroborate the viability of this frequency-specific detection method for ICH.

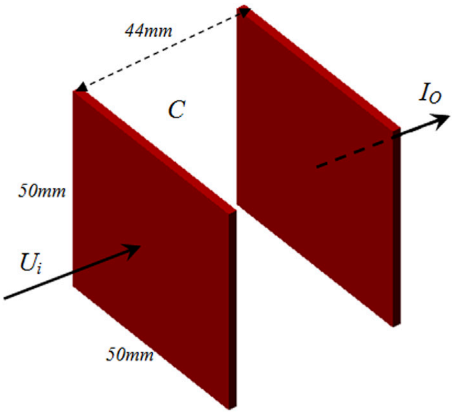


FIGURE 2
Schematic diagram of the parallel plate capacitor.

Methods and materials

Capacitance measurement system based on a parallel plate capacitor

The measurement system for the capacitance is shown in Figure 1, including a parallel plate capacitor and a 4294A impedance analyzer manufactured by Agilent. The capacitor employs two identical copper foils, each with a thickness of 0.1 mm, affixed to the exterior of a 3D-printed, cuboid-shaped barrel. Specific dimensions of the capacitor are presented in Figure 2. The electrodes measure 50 mm × 50 mm, with an inter-electrode distance of 44 mm and a barrel wall thickness of 2 mm. Connection wires are soldered centrally to each plate and subsequently interfaced with the measurement fixture of the impedance analyzer. The Agilent 4294A impedance analyzer operates within a test frequency range of 40 Hz–110 MHz and utilizes four-port measurement technology. This facilitates the assessment of impedance parameters such as resistance, inductance, and capacitance with a test accuracy up to 0.05%

[17]. A 16047E two-port fixture, specifically designed for dual-port component parameter measurement, is used as the measurement fixture. During the evaluation, the subject under investigation is positioned between the two plates of the parallel plate capacitor. According to Figure 2, let the capacitance between the two plates be *C*, then:

$$C = \frac{\varepsilon_r \cdot \varepsilon_0 \cdot S}{d} \tag{1}$$

where ε_r is the equivalent permittivity of the contents between the two plates, ε_0 is the vacuum permittivity, *S* and *d* are the surface area of the plate and the distance between the two plates respectively. Given constant plate area and distance, capacitance is directly proportional to the relative permittivity of the material situated between the plates. Thus, the capacitance values obtained from this parallel plate capacitor system serve as reliable metrics for gauging the permittivity of the examined material.

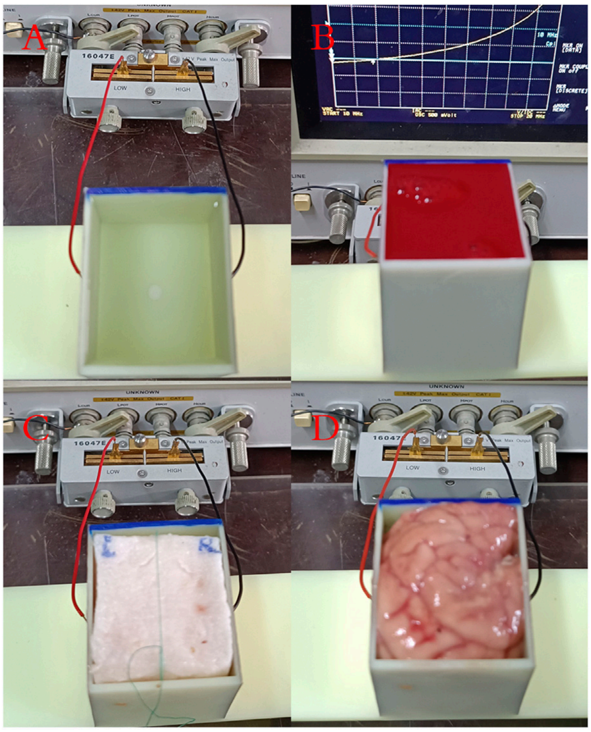


FIGURE 3
Measurement photos of four single-component substances. **(A)** Distilled water. **(B)** Sheep blood. **(C)** Pig fat. **(D)** Pig brain.

Capacitance frequency sweep measurements of single-component substance

The previously described parallel plate capacitor measurement system was employed to conduct frequency-dependent capacitance measurements on four single-component substances: distilled water, anticoagulant-treated sheep blood, pig fat, and pig brain. These substances were placed in the 3D-printed cuboid barrel shown in Figure 1, which has dimensions of 50 mm × 44 mm × 50 mm and a wall thickness of 2 mm. Photographic evidence of these measurements is provided in Figure 3. For the substances in liquid form, namely, water and blood, the barrel was filled to capacity. Solid pig fat was acquired commercially and cut into a cuboid measuring 46 mm × 40 mm × 50 mm before placement in the measurement barrel. Pig brain was gently stacked and compressed to fill the barrel. The fresh sheep blood was purchased from Alibaba's Taobao shopping app, and the seller has a license to sell animal blood. Fresh pig brain was purchased from Wal-Mart Supermarket in Chongqing. Fresh pig brain is a kind of fresh ingredient in China that can be freely bought and sold without ethical certification.

The impedance analyzer was configured with a frequency measurement range of 10 KHz–20 MHz. Measurement parameters were set to series capacitance, the data collection points were set at 150, and the accuracy level was maximized. For each substance, the measurement protocol involved initially filling the barrel and executing a single frequency sweep measurement with the impedance analyzer. Subsequently, the barrel was emptied and another measurement was taken. The data from these two steps were then subtracted to derive the change in capacitance (ΔC) at each frequency point. This procedure was repeated thrice for each substance to calculate the average ΔC at each frequency point. The four substances were evaluated in a sequential manner, adhering to the outlined protocol.

Capacitance frequency sweep measurements of mixed substances containing blood

Capacitance frequency sweep measurements were conducted on mixtures of blood with distilled water, pig fat, and pig brain, in accordance with the method delineated in Section 2. Photographic documentation of these experiments is presented in Figure 4. For the water-blood mixture (A), a needle tube with a 15 mm diameter (needle head removed) was positioned centrally within the distilled water. Sheep blood was then introduced into this tube until it reached the same height as the surrounding distilled water. For the fat-blood combination (B), a cylindrical hole with a 15 mm diameter was carved into the center of the cuboid pig fat sample and subsequently filled with sheep blood. Finally, in the brain-blood model (C), a 15 mm diameter needle tube was inserted centrally into the pig brain, and sheep blood was injected until the blood level equated that of the surrounding pig brain tissue. The measurement protocol remained consistent with that described in Section 2, thus ensuring uniformity in data acquisition across different experimental setups.

Results and discussions

Capacitance sweep measurement results of single-component substance

Figure 5 presents the results of raw ΔC measurements across three frequency ranges—10 KHz–1 MHz (A), 1 MHz–10 MHz (B),

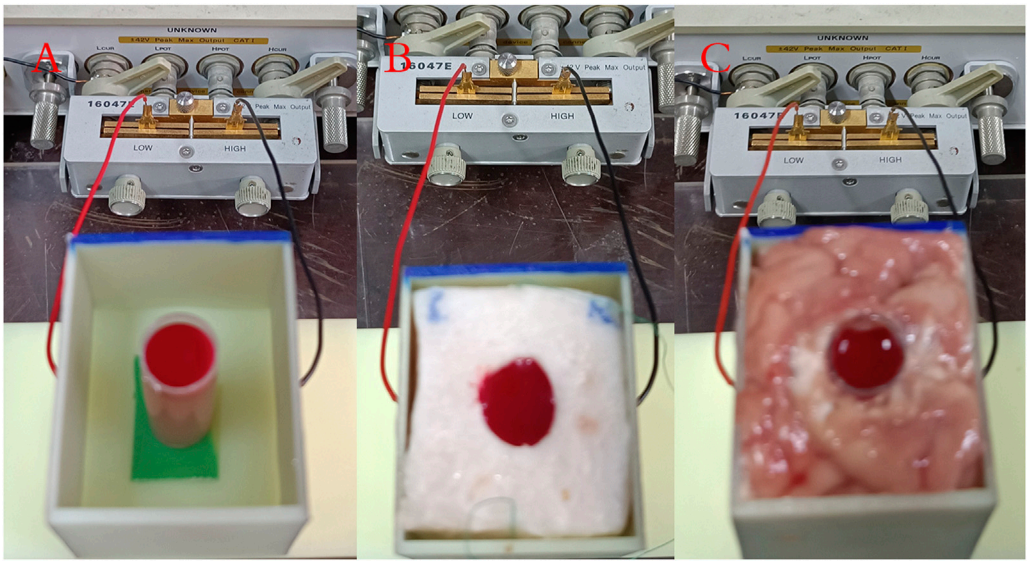


FIGURE 4
Measurement photos of three mixed blood substance models. **(A)** Water-blood mixture. **(B)** Fat-blood combination. **(C)** Brain-blood model.

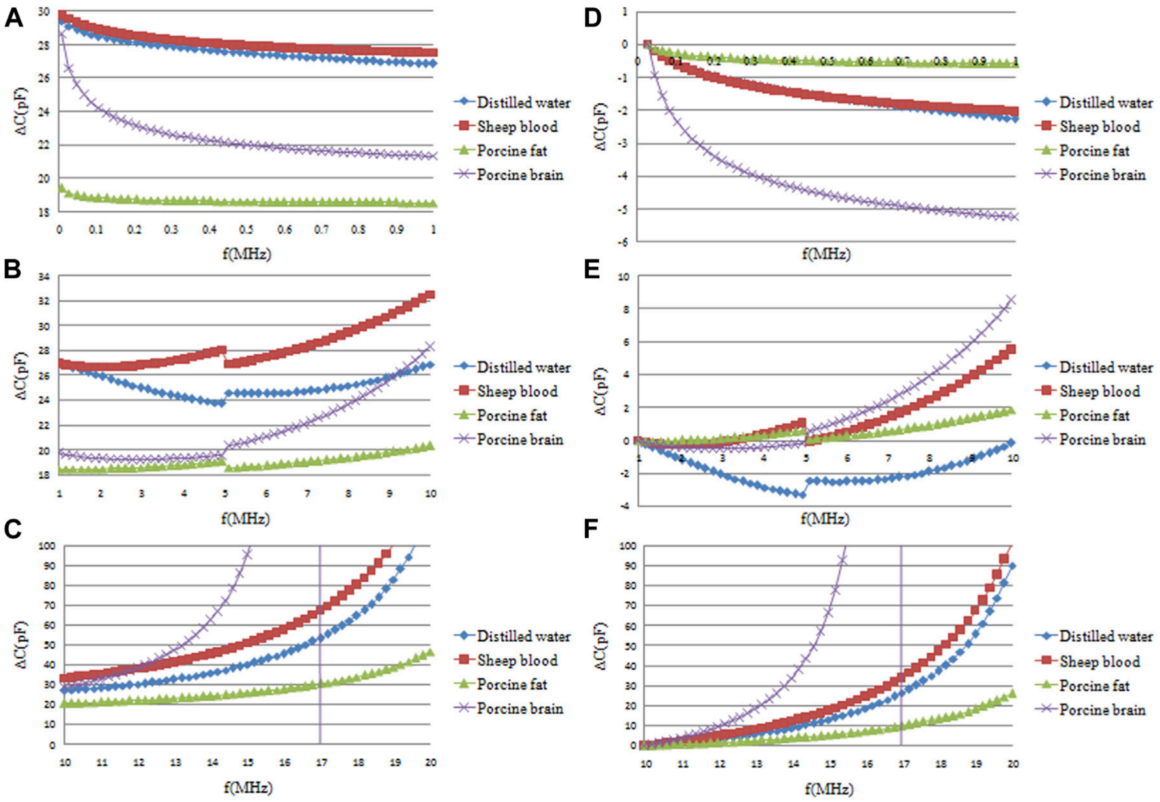


FIGURE 5
Frequency sweep measurement results of capacitance changes (ΔC) of four single-component substances: distilled water, sheep blood, pig fat and pig brain. **(A, B, C)** are the original ΔC in the three frequency ranges of 10 KHz–1 MHz, 1 MHz–10 MHz and 10 MHz–20 MHz, respectively. **(D, E, F)** are the normalized data of the original data in **(A, B, C)** relative to the ΔC data of the initial frequency.

and 10 MHz–20 MHz (C)—for distilled water, sheep blood, pig fat, and pig brain. Subfigures D, E, and F depict normalized data of the raw data in A, B, and C, scaled to the initial frequency data.

In the 10 KHz–1 MHz frequency range (A), the greatest ΔC is observed in sheep blood, followed by distilled water, with the smallest ΔC seen in pig fat. These results indicate a marginally higher permittivity for sheep blood relative to distilled water. This observation deviates from existing literature, which can be attributed primarily to the dilution effect of the added heparin sodium anticoagulant in the sheep blood used, bringing its permittivity closer to that of a liquid medium.

Moreover, both pig brain and pig fat exhibited smaller permittivity than water. This reduction is ascribed to the pre-experimental conditions—namely, the freezing and subsequent thawing of these tissues, leading to a significant incorporation of melted ice water, thereby diminishing the permittivity.

Analysis of A and D reveals that, within the frequency range of 10 KHz–1 MHz, the most pronounced decrement in ΔC is seen in pig brain, followed by distilled water and blood. Pig fat exhibits negligible frequency-dependent change, as indicated by an almost horizontal trend line. Consequently, while the difference in this frequency range between distilled water and blood is negligible, distinctions between blood, pig fat, and pig brain are readily discernible.

Analysis of data sets B and E indicates a discernible discontinuity at approximately 5 MHz across all four substances under investigation. Upon measuring the impedance of an unoccupied parallel plate capacitor, this irregularity was identified to emanate from a circuit resonance at around 5 MHz, resulting from the capacitance of the parallel plate capacitor and the circuit's parasitic inductance. In the 1 MHz–5 MHz frequency window, the ΔC of blood exhibits a gradual increase, whereas that of distilled water shows a decline; the changes for pig fat and pig brain are inconsequential in this range. Thus, differentiation between distilled water and blood becomes feasible in this specific frequency domain. The 5 MHz–10 MHz frequency span sees a general upward trend in ΔC for all substances. Pig brain registers the most significant increment, followed by blood. Distilled water's ΔC remains virtually static initially but starts ascending after surpassing the 8 MHz threshold. Data sets C and F provide insights into the 10 MHz–20 MHz range. In this frequency span, pig brain's ΔC skyrockets, eventually surpassing that of blood and inducing a new circuit resonance around 15 MHz. Across this frequency range, all substances exhibit a positive correlation between frequency and ΔC, albeit at varying magnitudes. Specifically, the values for blood and distilled water closely align, while pig fat shows the least amplitude variation. The data corroborate that the ΔC fluctuations differ among the four substances depending on the frequency range in question. Accordingly, frequency intervals manifesting the most substantial divergences may be leveraged for the effective discrimination of these substances.

However, the measurement results given in Figure 5 seems to be different from the data in Ref. [16]. According to the results in Ref. [16], the permittivity of blood and other brain tissues decrease gradually with increasing frequency. However, Figure 5 shows that the measurement results of blood, pig fat and pig brain all increase with increasing frequency in the 1 MHz–20 MHz frequency range. This is mainly because the measurement results in Figure 5 are the capacitance values, but the data in Ref. [16] are the permittivity values. Although it was mentioned earlier that the capacitance of a parallel plate capacitor is directly proportional to the dielectric constant of the material between the plates, this is conditional and only applies to low-frequency electrostatic fields. Therefore, in the frequency range of 10 kHz to 1 MHz, as shown in Figures 5A, D, the capacitance changes of the blood, pig fat, and pig brain—indeed decrease with increasing frequency. However, at higher frequencies, the impact of stray inductance and capacitance in the measuring circuit becomes more significant, and the capacitance measured by the impedance analyzer then shows certain discrepancies from the dielectric constant. Particularly when the frequency increases to a certain value, the entire measuring circuit undergoes resonance, as evidenced by the resonance occurring around 5 MHz in Figure 5B. After resonance, the direction of capacitance change measured also varies. Although at high frequencies, the capacitance measured by the impedance analyzer differs from the permittivity of the material being tested, this discrepancy is consistent across all materials tested. Therefore, the capacitance measured by the impedance analyzer and the parallel plate capacitor can be used to assess the differences in the permittivity of different materials.

Capacitance sweep measurement results of three kinds of mixed blood substances

The ΔC measurement results for both pure distilled water and its mixtures with blood are illustrated in Figure 6. Subsets A, B, and C depict raw ΔC measurement data across three frequency bands: 10 KHz–1 MHz, 1 MHz–10 MHz, and 10 MHz–20 MHz. D, E, and F represent the normalized versions of these data, scaled to the initial frequency ΔC. Data sets A and D indicate that both solutions display almost the same ΔC variations across the 10 KHz–1 MHz frequency range, rendering the identification of blood presence in distilled water challenging at these lower frequencies. In contrast, subsets B and E, which focus on the 1 MHz–5 MHz range, demonstrate a decreasing trend in ΔC as frequency escalates. Importantly, the magnitude of this decrement differs significantly between the two solutions. Specifically, the ΔC changes within the 1 MHz–5 MHz frequency range are quantified as 3.287 ± 0.11pF and 4.275 ± 0.18 pF for the two solutions, as shown in Figure 7. This implies that the blood-induced increment in the ΔC decrement is approximately 30% of that observed in pure distilled water, a differential that is easy to detect. Lastly, subsets C and F, which encompass the 10 MHz–20 MHz frequency range, reveal an increasing trend in ΔC for both solutions. However, the magnitude of this increase is largely comparable, resulting in closely aligned values. Upon a comprehensive analysis of the data, it can be concluded that the most favorable frequency domain for the differentiation of blood presence in distilled water lies between 1 MHz and 5 MHz.

There is a aberration in the results shown in Figure 6. The results in Figure 5 indicates that the ΔC measurement results of pure sheep blood is larger than that of pure distilled water. So the permittivity of a water-blood solution should be greater than that of distilled water alone theoretically, but the results presented in Figure 6 show the opposite. The reason lies in the methodology described in the article: as shown in Figure 4A, the water-blood solution was not created by

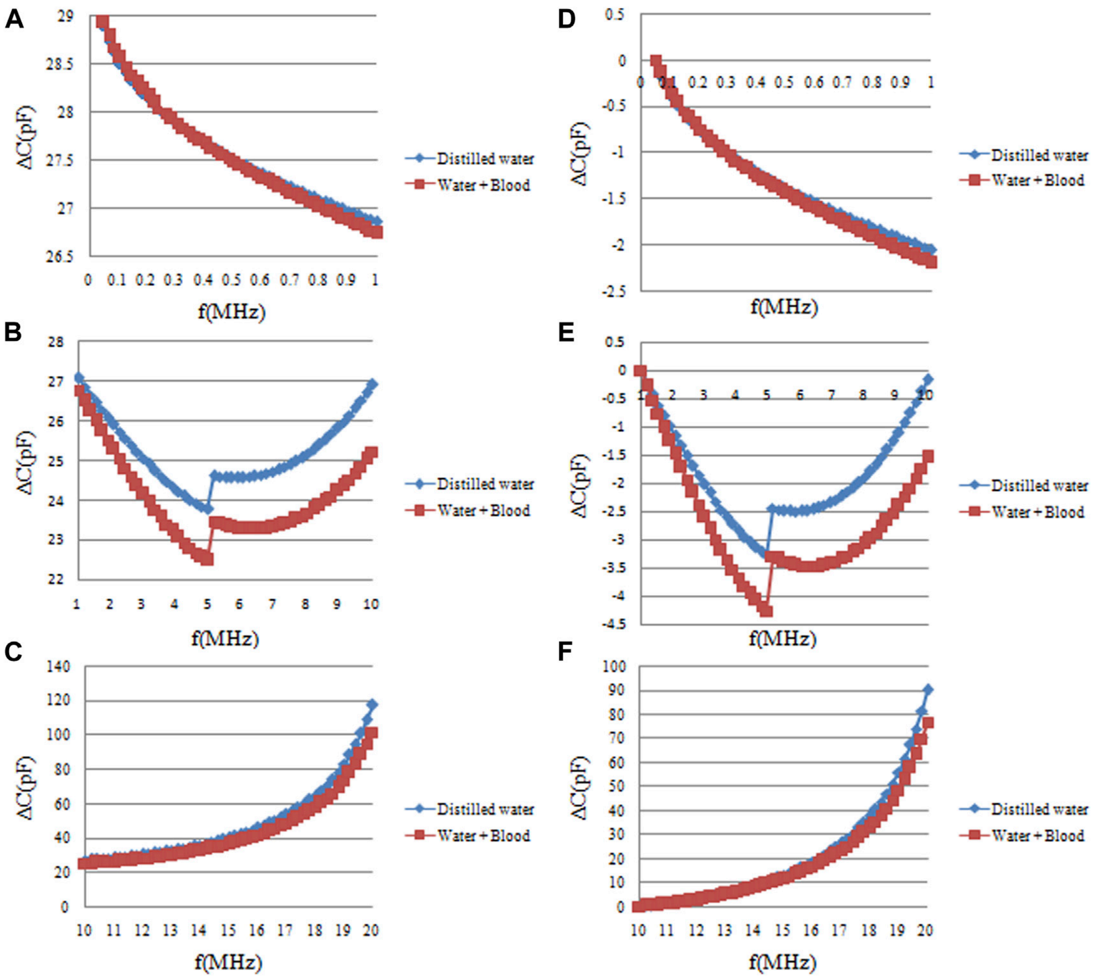


FIGURE 6
The frequency sweep measurement results of capacitance changes (ΔC) of pure distilled water and distilled water mixed blood. **(A, B, C)** are the original ΔC measurement data in the three frequency ranges of 10 KHz–1 MHz, 1 MHz–10 MHz, and 10 MHz–20 MHz, respectively. **(D, E, F)** are the normalized data of the original data in **(A, B, C)** relative to the data of the initial frequency of each frequency band.

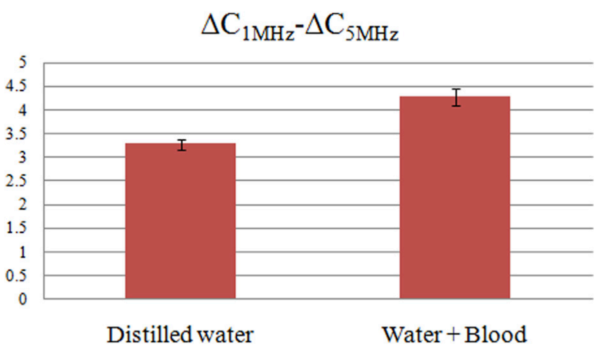


FIGURE 7
The ΔC differences between 1 MHz and 5 MHz for pure distilled water and distilled water mixed blood solution.

directly dissolving blood in water. Instead, the blood was first placed in a plastic tube, which was then placed in distilled water. Thus, there was a plastic film (1 mm thick) separating the blood from the distilled water. This plastic tube effectively isolates the distilled water and blood, acting as a capacitor. Although the dielectric constant of the plastic tube is very low, at higher frequencies, its capacitive effect becomes quite significant, thereby altering the overall capacitance value of the distilled water-blood solution.

The ΔC measurement results across various frequency bands for pure fat and fat-encapsulated blood are shown in Figure 8. Subsets A, B, and C are the raw ΔC measurement data in the frequency bands of 10 KHz–1 MHz, 1 MHz–10 MHz, and 10 MHz–20 MHz, respectively. D, E, and F represent the normalized data, scaled according to the initial frequency-specific ΔC. Subsets A and D reveal that in the 10 KHz–1 MHz frequency domain, both substances exhibit a declining trend in ΔC as frequency escalates. However, the magnitude of this decline differs substantially between the two, with a more pronounced decrement evident upon blood addition. Quantitative data, elaborated in Figure 9, establish ΔC changes for these substances in the 10 KHz–1 MHz range as 0.577 ± 0.041pF and 0.670 ± 0.036pF, respectively. The blood-induced augmentation in ΔC decline constitutes 16% of that of the pure fat, a differential that is discernible. In the 1 MHz–10 MHz

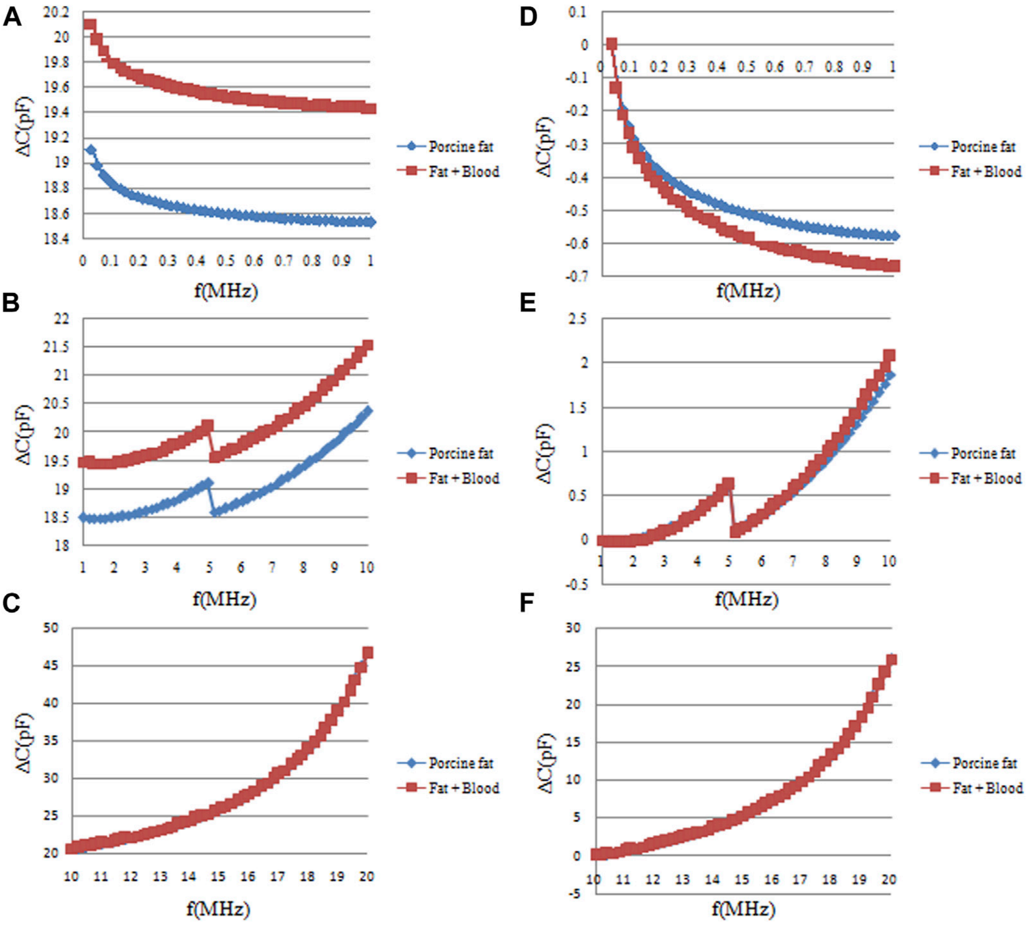


FIGURE 8
The frequency sweep measurement results of capacitance changes (ΔC) of pure fat and fat-wrapped blood. **(A, B, C)** are the original ΔC measurement data in the three frequency ranges of 10 KHz–1 MHz, 1 MHz–10 MHz, and 10 MHz–20 MHz, respectively. **(D, E, F)** are the normalized data of original data in **(A, B, C)** relative to the data of the initial frequency of each frequency band.

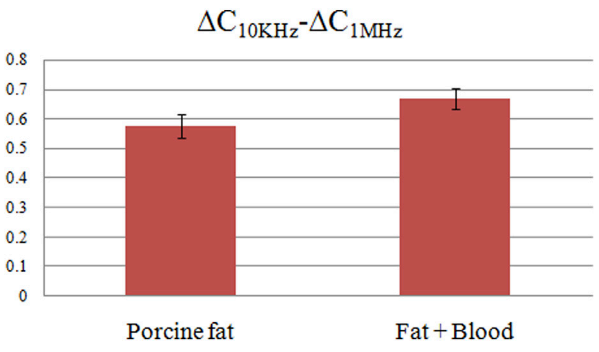


FIGURE 9
The ΔC differences between 10 KHz and 1 MHz for pure pig fat and pig fat wrapped blood.

frequency band, represented by subsets B and E, the differential in ΔC changes between the substances is marginal. While a larger ΔC change in the fat-encapsulated blood model is observed within the 5 MHz–10 MHz subset, the differential is not markedly pronounced. Subsets C and F, capturing data in the 10 MHz–20 MHz band, indicate almost identical ΔC changes between the substances, negating the possibility of discerning blood presence in fat at these frequencies. A holistic analysis suggests that the 10 KHz–1 MHz frequency range is the most efficacious for detecting the presence of blood in pig fat through frequency differential analysis.

The ΔC measurement results for both pure pig brain and blood-infused pig brain are shown in Figure 10. Subsets A, B, and C represent the raw ΔC measurement data across three distinct frequency bands: 10 KHz–1 MHz, 1 MHz–10 MHz, and 10 MHz–20 MHz. Subsets D, E, and F provide normalized data corresponding to the raw measurements in A, B, and C, scaled relative to the initial frequency data. Within the 10 KHz–1 MHz range, subsets A and D indicate a frequency-dependent decrease in ΔC for both pure pig brain and blood-infused pig brain. However, the magnitude of this decline differs between the two, with a more modest reduction observed post-blood infusion. This observation is corroborated by frequency sweep data of pure pig brain and pure blood depicted in Figure 5; within the same frequency range, the rate of ΔC decline in pure pig brain substantially exceeds that of pure blood. Consequently, the introduction of blood into the pig brain reduces the overall decline in ΔC, yet this alteration does not

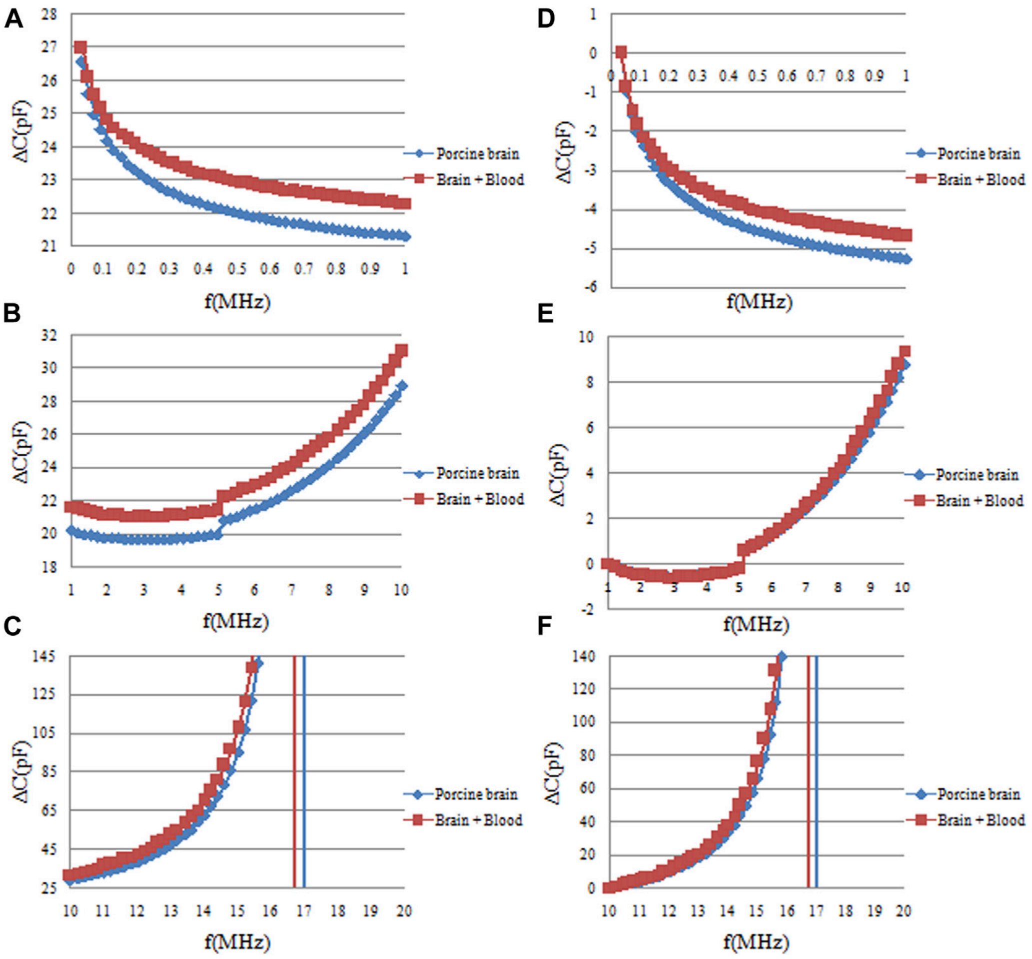


FIGURE 10
The frequency sweep measurement results of capacitance changes (ΔC) of pure pig brain and pig brain wrapped blood. **(A, B, C)** are the original ΔC measurement data in the three frequency ranges of 10 KHz–1 MHz, 1 MHz–10MHz, and 10 MHz–20 MHz, respectively. **(D, E, F)** are the normalized data of the original data in **(A, B, C)** relative to the data of the initial frequency of each frequency band.

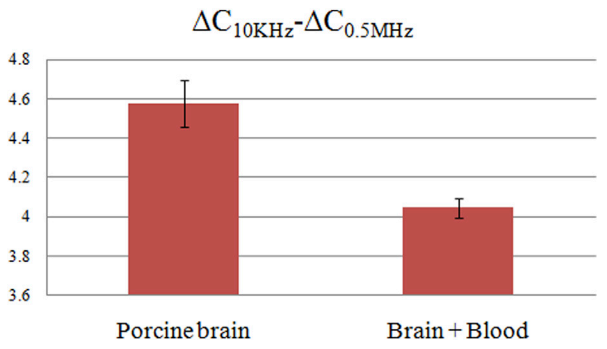


FIGURE 11
The ΔC differences between 10 KHz and 0.5 MHz for pure pig brain and pig brain wrapped blood.

impede its detect ability. Figure 10 subsets A and D further reveal relatively large difference of ΔC decline between the two materials in the 10 KHz–0.5 MHz range, but negligible difference in the 0.5 MHz–1 MHz range. These quantitative differences are documented in Figure 11, with values of 4.576 ± 0.12pF and 4.048 ± 0.05pF, respectively. Therefore, the reduced ΔC decline due to the infusion of blood represents 11.5% of the ΔC decline observed in pure pig brain, a difference that remains within detectable limits. Subsets B and E indicate almost the same ΔC fluctuations in the frequency band of 1 MHz–10 MHz for two materials. Subsets C and F delineate that within the frequency range of 10 MHz–15 MHz, the divergence in ΔC fluctuations between pure pig brain and blood-infused pig brain remains minimal. It should be noted that at approximately 16 MHz, the measurement circuit exhibited resonance, a factor which necessitates consideration in future experimental setups. Continuing from the previous data analysis, Upon aggregating the data, it is concluded that the most effective frequency domain for ascertaining the presence of blood in pig brain is between 10 KHz and 0.5 MHz. However, within this specified range, the decrease in ΔC decline due to blood infusion amounts to only 11.5% of the ΔC decline observed in the pure pig brain. This percentage is markedly lower than the 16% registered in pig fat and the 30% observed in distilled water, as per previous studies. Hence, the task of detecting blood in pig brain via frequency-dependent capacitance changes proves more challenging compared to the other models.

Conclusion

At present, whether ECT or other electrical imaging, ICH can only be detected by time-difference measurements. Such an approach necessitates baseline, non-hemorrhagic measurements data from the patient, which is often impracticable to acquire in a clinical setting. To address this shortfall, the present study advocates for the utilization of frequency-dependent permittivity variations as an alternative modality for the absolute identification of ICH. This strategy leverages the distinct dispersion properties of the permittivity in brain tissue, along with the spectral disparity in permittivity between blood and other cerebral tissues. Notably, this frequency-based approach obviates the need for non-hemorrhagic reference data. It mandates only the acquisition of measurement data at specific frequency points post-hemorrhage, thereby rectifying the limitations inherent to time-difference imaging techniques. Initial steps in this method involve identifying two critical frequency points at which the permittivity frequency difference of blood and surrounding brain tissues exhibit maximum variations. Subsequently, the differential permittivity at these frequencies are employed to detect ICH. In the scope of this paper, the capacitance variations of four individual substances and three composite materials were quantified using an impedance analyzer in conjunction with a parallel plate capacitor.

The study confirms that distinct frequency bands can effectively discriminate between single-component and mixed substances based on variations in capacitance with frequency. Comparative analysis reveals that the frequency range between 1 MHz and 5 MHz is most efficacious for the detection of blood in distilled water. For the identification of blood in pig fat, the optimal frequency range is between 10 KHz and 1 MHz. When isolating blood within pig brain tissue, the frequency range of 10 KHz to 0.5 MHz demonstrates optimal results. Thus, the study establishes that the absolute detection of ICH is feasible through frequency-differential measurements of the permittivity. Nonetheless, this research is not without limitations. The sheep blood used in the experiments of this paper was diluted with sodium heparin anticoagulant, resulting in its permittivity being significantly lower than *in vivo* blood, only slightly higher than that of water. This was proven in our preliminary experiments. The use of such treated blood was a necessary compromise, as fresh blood coagulates very quickly. The pig fat and pig brain used in the experiments were also previously frozen in a refrigerator. Before the experiments, they were thawed out, and during this process, melting ice water was also mixed in. Therefore, the permittivity of these *ex vivo* pig fat and pig brain tissues differs greatly from that of *in vivo* tissues. The primary objective of this paper was to verify the feasibility of using the frequency differences in permittivity to detect cerebral hemorrhage. Therefore, we initially conducted verification through experiments with *ex vivo* tissues, with plans to validate further through *in vivo* animal experiments in the later stages. Additionally, there is a divergence in the capacitance sweep measurements obtained in this study compared to those reported in existing literature. This discrepancy can be largely attributed to the measurement circuitry. The parallel plate capacitor, along with parasitic inductance within the circuit, manifests varying impedances at different frequencies, leading to resonance at specific frequencies.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

SP: Investigation, Methodology, Writing–original draft. XW: Data curation, Formal Analysis, Investigation, Methodology, Resources, Software, Writing–original draft. GJ: Conceptualization, Methodology, Supervision, Writing–review and editing. FW: Conceptualization, Resources, Supervision, Visualization, Writing–review and editing. JZ: Data curation, Supervision, Validation, Writing–review and editing. XZ: Supervision, Validation, Visualization, Writing–review and editing. NL: Conceptualization, Funding acquisition, Methodology, Project administration, Writing–review and editing. RX: Funding acquisition, Methodology, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the 2023 Chongqing Technology Innovation and Application Development Special Science and Technology Assistance to Tibet and Xinjiang Project (CSTB2023TIAD-KPX0001), and 2023 Natural Science Foundation of Xizang Autonomous Region [XZ2023ZR-ZY67 (Z)].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Neurology CSO, Society CS. *Chinese guidelines for diagnosis and treatment of cerebral venous thrombosis 2019* (2020). doi:10.3760/cma.j.cn113694-20200225-00113

2. Thrift AG, Thayabaranathan T, Howard G, Howard VJ, Rothwell PM, Feigin VL, et al. Global stroke statistics[J]. *Int J Stroke: official J Int Stroke Soc* (2017) 12:13–32. doi:10.1177/1747493016676285

3. Wang L, Liu J, Yang G, Peng B, Wang Y. The prevention and treatment of stroke in China is still facing great challenges—summary of China Stroke Prevention report 2018. *Chin Circ J* (2019) 34(02):6–20.

4. Cheng J, Wang W, Xu J, Yin L, Liu Y, Wu J. Trends in Stroke Mortality Rate-China, 2004–2019[J]. *Chinese Center for Disease Control and Prevention Weekly Report* (2022) (024):004.

5. Bodenstein M, David M, Markstaller K. Principles of electrical impedance tomography and its clinical application. *Crit Care Med* (2010) 37(2):713–24. doi:10.1097/ccm.0b013e3181958d2f

6. Griffiths H. Magnetic induction tomography. *Meas Sci Techn* (2001) 12:1126–31. doi:10.1088/0957-0233/12/8/319

7. Sasaki K, Wake K, Watanabe S Development of best fit Cole-Cole parameters for measurement data from biological tissues and organs between 1 MHz and 20 GHz. *Radio ence* (2015) 49(7):459–72. doi:10.1002/2013RS005345

8. Warsito W, Marashdeh Q, Fan LS. Electrical capacitance volume tomography. *IEEE Sensors J* (2007) 7(4):525–35. doi:10.1109/jsen.2007.891952

9. Wang H, Yang W. Application of electrical capacitance tomography in pharmaceutical fluidised beds – a review. *Chem Eng Sci* (2020) 231:116236. doi:10.1016/j.ces.2020.116236

10. Bai Z, Li H, Chen J, Zhuang W, Li G, Chen M, et al. Research on the measurement of intracranial hemorrhage in rabbits by a parallel-plate capacitor. *PeerJ* (2021) 9(99): e10583. doi:10.7717/peerj.10583

11. Xu R, Zhuang W, Bai Z, Wang F, Jin G, Liu N, et al. A pilot study on intracerebral hemorrhage imaging based on electrical capacitance tomography. *Front Phys* (2023) 3(11). doi:10.3389/fphy.2023.1165727

12. Chen Q, Liu R, Wang C, Liu RJMS Real-time *in vivo* magnetic induction tomography in rabbits: a feasibility study. *Meas Sci Technol* (2020) 32(3):035402. doi:10.1088/1361-6501/abc579

13. Holder D, Saulnier G. *Electrical impedance tomography methods history and applications series in medical Physics and biomedical engineering by david holder*. England, UK: Routledge (2020).

14. Watson SR, Williams JW, Gough A, Griffiths H. A magnetic induction tomography system for samples with conductivities below 10 S m^{-1}. *Meas Sci Techn* (2008) 19(4):045501. doi:10.1088/0957-0233/19/4/045501

15. Malikov V, Tihonsky NN, Kozlova DV, Ishkov AV. Computer-aided study of the impedance dispersion of biological tissues. *J Phys Conf Ser* (2021) 2142:012012. doi:10.1088/1742-6596/2142/1/012012

16. Gabriel S, Lau RW, Gabriel S, Lau RW. The dielectric properties of biological tissues: II. Measurements in the frequency range 10 Hz to 20 GHz. *Phys Med Biol* (1996) 41:2251–69. doi:10.1088/0031-9155/41/11/002

17. Mohamed M. A method for modeling a common-mode impedance for the AC motor. *Elektrotehniski Vestnik/Electrotechnical Rev* (2017) 84(5).

Frontiers in Physics

TYPE Original Research
PUBLISHED 24 January 2024
DOI 10.3389/fphy.2024.1335494

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Xin Huang,
Chongqing University of Posts and Telecommunications, China
Yi Li,
Shanghai Maritime University, China
Yishun Li,
Tongji University, China
Rencan Nie,
Yunnan University, China

*CORRESPONDENCE
Shanchuan Yu,
✉ yushanchuan@cmhk.com
Jun Wang,
✉ 277681785@qq.com
Lijia Fu,
✉ fulijia@cmhk.com

RECEIVED 09 November 2023
ACCEPTED 02 January 2024
PUBLISHED 24 January 2024

CITATION
Yang X, Yu S, Wang J, Chen H, Huang Y, Luo Z and Fu L (2024), Application of a digital twin for highway tunnels based on multi-sensor and information fusion.
Front. Phys. 12:1335494.
doi: 10.3389/fphy.2024.1335494

Application of a digital twin for highway tunnels based on multi-sensor and information fusion

Xun Yang[1,2], Shanchuan Yu[1,2]*, Jun Wang[3]*, Hong Chen[4], Yonggang Huang[5], Zhongbin Luo[1,2] and Lijia Fu[1,2]*

[1]China Merchants Chongqing Communications Research and Design Institute Co., Ltd., Chongqing, China, [2]Research and Development Center of Transport Industry of Self-driving Technology, Chongqing, China, [3]Gansu Provincial Highway Aviation Tourism Investment Group Co., Ltd., Gansu, China, [4]Shandong Hi-speed Company Limited, Shandong, China, [5]Guangxi Computing Center Co., Ltd., Guangxi, China

Due to the harsh environment of highway tunnels and frequent breakdowns of various detection sensors and surveillance devices, the operational management of highway tunnels lacks effective data support. This paper analyzes the characteristics of operational surveillance data in highway tunnels. It proposes a multimodal information fusion method based on CNN–LSTM–attention and designs and develops a digital twin for highway tunnel operations. The system addresses issues such as insufficient development and coordination of the technical architecture of operation control systems, weak information service capabilities, and insufficient data application capabilities. The system also lacks intelligent decision-making and control capabilities. The developed system achieves closed-loop management of "accurate perception–risk assessment–decision warning–emergency management" for highway tunnel operations based on data-driven approaches. The engineering demonstration application underscores the system's capacity to enhance tunnel traffic safety, diminish tunnel management costs, and elevate tunnel driving comfort.

KEYWORDS

highway tunnel, operational surveillance, information fusion, deep learning, digital twin

1 Introduction

Tunnels are vital infrastructures on national highway networks. As of the end of 2021, there were 23,268 road tunnels in China, with a total length of 24,698.9 km, an increase of 1,952 tunnels and 2,699.6 km, of which there were 1,599 special long tunnels with a length of 7,170.8 km and 6,211 long tunnels with a length of 10,844.3 km [1]. To supervise the operation of highway tunnels, tunnels are equipped with electromechanical systems which consist of communication and monitoring facilities. During the construction and operation of highway tunnel electromechanical systems, a large amount of experience has been accumulated and complete engineering technical specifications and related product standards have been formed, laying a good foundation for realizing intelligent management and control of highway tunnel operations. With the large-scale construction and operation of Chinese tunnels, especially the increasingly mature digital tunnel monitoring technology, most highway tunnels currently use area control and switch ring network control transmission technology, which has realized equipment control and data monitoring management inside the tunnel, and the technology is mature and stable [2].

As early as the 1960s, countries such as the United States, Germany, and Japan began researching about the mechanical and electrical control systems of highway tunnels. In the mid-1970s, Japan's OMRON Corporation studied a controller link network control technology that could form a redundant token ring network with autonomous control capabilities. In the mid-1980s, Germany's Phoenix Contact Corporation researched an INTERBUS fieldbus technology that is suitable for network technology in the automatic control field. Based on the INTERBUS technology, a highway tunnel operation control system can perform distributed monitoring of the mechanical and electrical equipment in single-tube two-way highways or dual-tube one-way highways. Each tunnel can have an independent centralized monitoring center. In the mid-1990s, the United States' Echelon Corporation introduced a LonWorks distributed highway tunnel monitoring system, which was developed based on Echelon's OpenLNS Server software platform and possessed characteristics such as distribution, openness, interoperability, and adaptability. These research results have been widely recognized by the industry and have become mainstream standards for control networks, widely applied in the mechanical and electrical control of highways and tunnels worldwide [3–5].

Since the beginning of the 21st century, countries have gradually improved the theoretical and technical aspects of highway tunnel operation safety monitoring. The United States, Germany, and other countries have formulated technical standards and traffic control regulations for highway tunnel operation management systems, while continuously researching and developing updated highway monitoring facilities and software systems with the application of new technologies [6–9]. Research on tunnel operation management technology in China began in the 1980s. Researchers gradually conducted research on tunnel structural design, safety construction monitoring, and operational safety management. The Shanghai Yan'an East Road Tunnel and the Shenzhen Wutongshan Tunnel were the earliest to introduce foreign tunnel monitoring systems for application. The Zhongliangshan Tunnel and the Jinyunshan Tunnel of the Chongqing–Guizhou Highway also successively introduced advanced equipment from abroad. With the accumulation of experience in the construction and operation of highway electromechanical systems, a series of national standards and industry specifications related to highway tunnel traffic engineering design were formed. A number of successful cases have been established, such as the Hong Kong–Zhuhai–Macao Bridge long-span immersed tunnel monitoring and management platform, the Qinling–Zhongnanshan Mountain long tunnel monitoring and management platform, the Shandong Province Jilai Highway tunnel monitoring and management platform, and the CMCT's new generation of intelligent highway tunnel management platform, which has laid a good foundation for achieving intelligent management and control of highway network traffic operation. In recent years, with the continuous advancement of China's construction of a transportation power, provinces such as Shandong, Zhejiang, Jiangsu, Gansu, Henan, and Yunnan have successively issued local smart highway construction guidelines [10–15]. Industry software and hardware leaders such as Hikvision, Huawei, and Wanji also successively released smart highway overall solutions [16, 17]. As tunnels are key nodes of highways, higher requirements have been put forward for intelligent tunnel operation management systems [18–21].

The data in highway tunnels exhibit characteristics such as multiple sources and heterogeneity. Various pieces of electromechanical equipment inside the tunnel constantly generate a massive amount of data, including real-time monitoring data such as video, environmental conditions, and traffic data collected by vehicle detectors (e.g., traffic flow, speed, and occupancy). The equipment control parameters include ventilation-, lighting- and traffic-related ones. Due to the harsh operating environment in the tunnel, the operational data detection equipment, such as CO/VI detectors, wind speed and direction detectors, brightness detectors, and vehicle detectors, experiences frequent failures, leading to distorted monitoring data. This hampers the provision of data-based guidance for the precise control of ventilation, lighting, traffic, and guidance. Currently, research and application of data mining in highway tunnel monitoring systems are almost nonexistent, with few relevant research achievements. The road condition perception system on the Yanqing to Chongli highway employs LiDAR-vision fusion for road information acquisition [22]. However, this approach is hampered by the absence of an integrated data fusion process, leading to asynchronous radar and video data collection. This limitation results in a restricted range of data on individual vehicle operations. In contrast, the Jinghu Jilai Highway tunnels employ a more advanced traffic data collection system, integrating technologies such as laser radar, millimeter-wave radar, and panoramic cameras. The practical outcomes of this system reveal that traffic data collection, when based on a multi-sensor fusion approach, significantly surpasses the accuracy of target recognition achievable with a single sensor. Furthermore, the implementation of a bidirectional optimal estimation algorithm, built upon the fused data, enhances the reliability of traffic flow data collection.

In its role as a facilitator of intelligent highway tunnel operation, the current mechanical and electrical system still has the following problems in technology architecture, functional design, and operation management: the development of the system's technical architecture is not enough, the coordination is insufficient, and it is difficult to adapt to the edge-cloud architecture; the system's functions have relatively single information exchange modes and weak precision and timely information service functions, and the data intelligent aggregation, analysis, and application capabilities need to be improved urgently. Furthermore, the establishment of intelligent decision control capability remains a pending endeavor. This article initiates its exploration from the feature analysis of monitoring data pertinent to highway tunnel operations in Section 2. Aiming to significantly elevate the accuracy of highway tunnel operational status assessments, this is achieved through the implementation of a multimodal information fusion method grounded in CNN–LSTM–attention in Section 3. It designs and develops a digital twin system for highway tunnel operations in Section 4, realizing a closed-loop management of "precise perception–risk judgment–decision warning–emergency control" for highway tunnel based on data-driven approaches.

TABLE 1 Collection methods and requirements of highway tunnel operation data.

Data	Collection method	Sensor installation requirement
Speed data	Vehicle detection sensor	1. When using induction coil detection sensors, the spacing should be arranged between 300–750 m
Traffic volume data		2. Vehicle detection sensors (radar, microwave, LiDAR fusion sensors, etc.) should be set up to prevent other equipment or objects from blocking
Percentage of lane occupancy data		
Video data	Video image sensor	1. The video image sensor outside the tunnel should be set at the entrance and exit of the tunnel between 100–250 m
		2. For the video image sensor inside the tunnel, a spacing of 100–200 m should be used at a distance of 2–5 m from the entrance, and the recommended setting is 120–150 m
Carbon monoxide data	Carbon monoxide detection sensor	1. For tunnels with jet fans for longitudinal ventilation, they should be set up in the middle, at the bends, and a distance of 100–150 m from the exit
Visibility data	Visibility detection sensor	2. For tunnels with vertical and inclined shaft ventilation, they should be set up 30 m in front of the exhaust port
		3. The detection sensor is installed on the outer side wall bracket of the tunnel, with a height of 2.5–3 m from the maintenance road
Nitrogen dioxide data	Nitrogen dioxide detection sensor	The detection sensor is installed on the outer side wall bracket of the tunnel, with a height of 2.5–3 m from the maintenance road
Light intensity data outside the tunnel	Light intensity detection sensor	Light intensity detection sensor outside the tunnel is installed outside the tunnel, a distance of one parking line of sight (100–200 m) from the tunnel entrance
Light intensity data inside the tunnel	Light intensity detection sensor	Light intensity detection sensor inside the tunnel is installed inside the tunnel, 20–25 m away from the tunnel entrance
Wind speed and direction data	Wind speed and direction detection sensor	1. For tunnels with jet fans for longitudinal ventilation, they should be set up at the bends and a distance of 100–150 m from the exit
		2. For tunnels with vertical and inclined shaft ventilation, they should be set up 30 m in front of and behind the exhaust and supply air outlets
		3. The detection sensor is installed on the outer side wall bracket of the tunnel, with a height of 2.5–3 m from the maintenance road or installed on the nails on both sides of the inside and outside of the tunnel, and the two probes make an angle of 30°–60° with the longitudinal center line of the tunnel, preferably 45°, and cannot encroach on the building clearance
Traffic event data	Event monitoring sensor	1. Setting principles refer to video image sensors
		2. It is recommended to repurpose existing video image sensors. If using fusion perception devices, they can replace the original video image sensors

2 Acquisition and feature analysis of operational monitoring data for highway tunnels

2.1 Operational data collection of highway tunnels

The discrimination and control of the operating status of highway tunnels mainly rely on the monitoring facilities on site. Monitoring facilities generally include monitoring, control, and induction facilities. Monitoring facilities typically consist of vehicle detection facilities, environmental detection facilities, road anomaly detection facilities, video surveillance facilities, and alarm facilities [23–25]. Control and induction facilities include emergency call facilities, information release and control facilities, and local control facilities. The monitoring settings related to highway tunnel operation data collection mainly include vehicle detection sensors, light intensity detection sensors, carbon monoxide/visibility detection sensors, wind speed and direction detection sensors, and video image sensors, and the collection content, methods, and installation requirements are shown in Table 1.

TABLE 2 Characteristics of highway tunnel operation data.

Tunnel environment	Direction data	Data units	Data range	Data accuracy	Data transmission cycle (min)
Ventilation environment monitoring	Visibility data	m-1	0–0.0015 m-1	± 0.0002 m-1	5–10
	Carbon monoxide data	10–6 (ppm)	0–300 × 10^{-6} (0–300 ppm)	± 2 × 10^{-6} (± 2 ppm)	5–10
	Wind speed and direction data	m/s	−20 ~ +20 m/s	± 0.2 m/s	5–10
	Nitrogen dioxide data	10–6 (ppm)	0–10 cm3/m3	± 5% indicated value	5–10
Lighting environment monitoring	Light intensity data outside the tunnel	cd/m2	1–6500cd/m2	± 3% indicated value	5–10
	Light intensity data inside the tunnel	lx	1–1000lx	± 3% indicated value	5–10
Traffic environment monitoring	Speed data	Km/h	5–2000 km/h	Accuracy ≥ 85%	5–10
	Traffic volume data	Vehicle/h	-	Accuracy ≥ 85%	5–10
	Percentage of lane occupancy data	Vehicle/km	-	Accuracy ≥ 85%	5–10

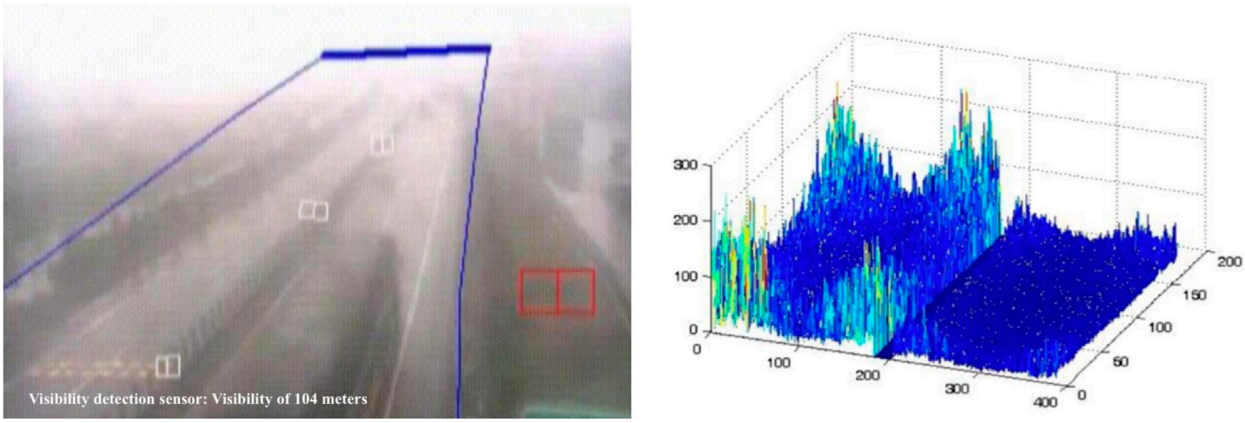


FIGURE 1
Image-based visibility detection and recognition.

2.2 Characteristic analysis of highway tunnel operation data

In the process of managing the operation of highway tunnels, traffic and environmental conditions are the focus of attention for managers. The spatiotemporal variation of traffic volume and composition, as well as vehicle speed, can determine whether there are traffic safety hazards in the tunnel and whether the luminaires meet the requirements of traffic safety [26–28]. This paper selects data from ventilation, lighting, and traffic detection sensors for analysis. The data characteristics are shown in Table 2, and the data correlations are shown in Eqs 1–4. The data collected by the vehicle detection sensors and visibility detection sensors can be calibrated through video images [29–33], as shown in Figure 1.

$$Q_{VI} = \frac{1}{3.6*10^6} * q_{VI} * f_{a(VI)} * f_d * f_{h(VI)} * f_{iv(VI)} * L * \sum\nolimits_{m=1}^{n_D} \left(N_m * f_{m(VI)}\right). \tag{1}$$

This equation includes the following variables: Q_{VI} is the smoke emission amount of the tunnel; q_{VI} is the benchmark smoke emission amount for the target year, which can be calculated based on the specifications; $f_{a(VI)}$ is the coefficient of vehicle condition considering smoke, which is determined according to the specifications; f_d is the vehicle density coefficient, which is determined according to the specifications; $f_{h(VI)}$ is the altitude coefficient considering smoke, which is determined according to the specifications; $f_{iv(VI)}$ is the longitudinal slope-speed coefficient considering smoke, which is determined according to the specifications; L is the length of the tunnel; $f_{m(VI)}$ is the diesel vehicle type coefficient considering smoke; n_D is the number of diesel vehicle type categories; and N_m is the traffic volume of the corresponding vehicle type, which is determined according to the specifications.

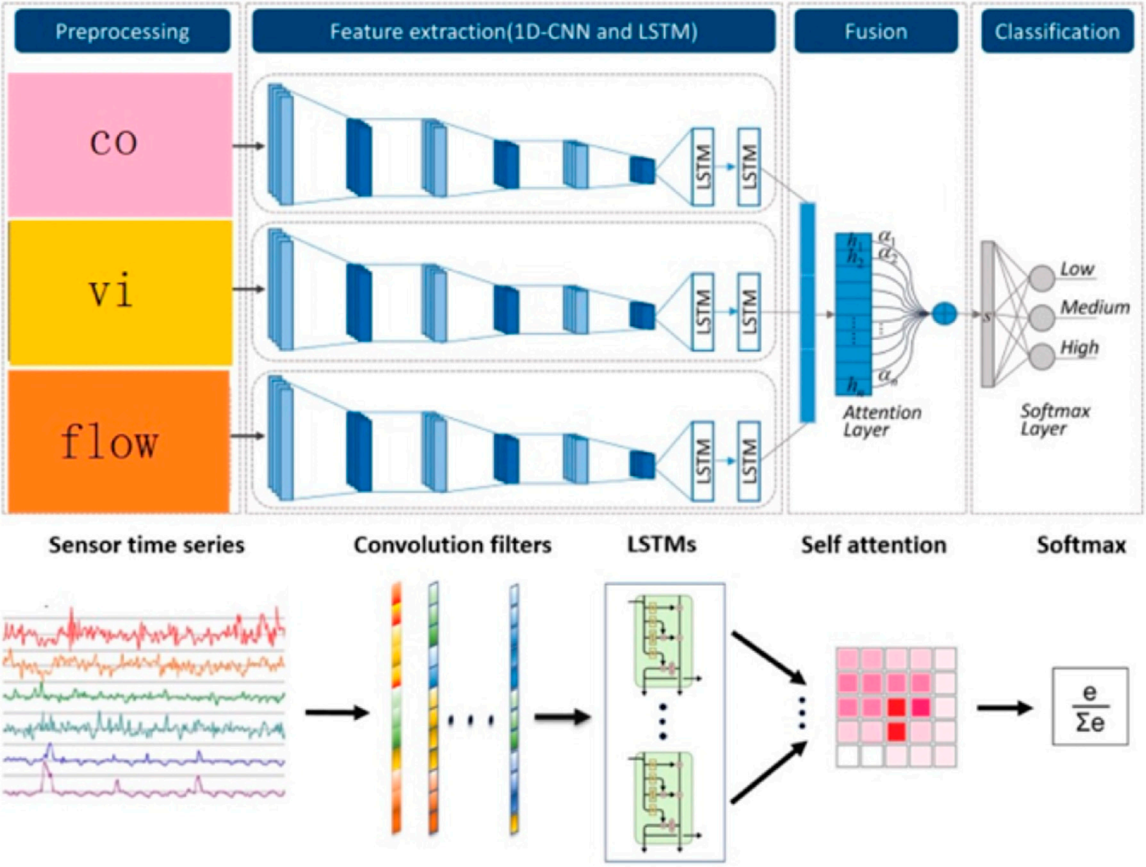


FIGURE 2
Multimodal fusion network architecture of tunnel operation monitoring data.

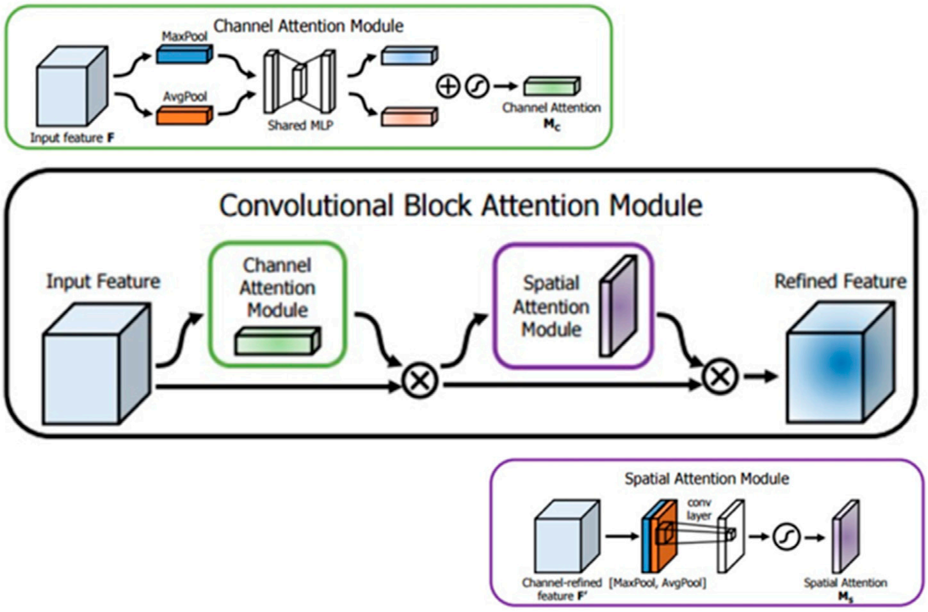


FIGURE 3
CBAM attention mechanism structure diagram.

TABLE 3 1D CNN–LSTM model parameters.

Layer	Parameter
Convolutional layers	Filter = 20, kernel size = (10.1), and stride = 1
Max pooling layer + dropout (0.15)	Pool size = (2.1) and stride = 2
Convolutional layers	Filter = 40, kernel size = (5.1), and stride = 1
Max pooling layer + dropout (0.15)	Pool size = (2.1) and stride = 2
Convolutional layers	Filter = 80, kernel size = (3.1), and stride = 1
Max pooling layer + dropout (0.15)	Pool size = (2.1) and stride = 2
LSTM	Hidden size = 64

$$Q_{CO} = \frac{1}{3.6 * 10^6} * q_{CO} * f_a * f_d * f_h * f_{iv} * L * \sum_{m=1}^{n_D} (N_m * f_m). \quad (2)$$

This equation describes the calculation of carbon monoxide (CO) emissions in a highway tunnel. Q_{CO} represents the amount of CO emissions, while q_{CO} is the baseline emissions rate for the target year, which can be calculated based on relevant specifications. The coefficients f_a, f_d, f_h, and f_{iv} represent the effects of vehicle condition, traffic density, altitude, and slope-velocity on CO emissions, respectively, and they are obtained according to specifications. L is the length of the tunnel. f_m is the coefficient for diesel vehicle type considering CO emissions, n_D is the number of diesel vehicle types, and N_m is the traffic volume for the corresponding vehicle type, which are all determined based on specifications.

$$L_{th1} = k * L_{20}(S). \quad (3)$$

This equation includes the following variables: L_{th1} represents the brightness of the TH1 section at the tunnel entrance; L_{th1} represents the brightness of the TH2 section at the tunnel entrance; k is the reduction coefficient of the entrance section brightness, which is obtained by consulting the specifications based on traffic volume data; and $L_{20}(S)$ represents the brightness outside the tunnel.

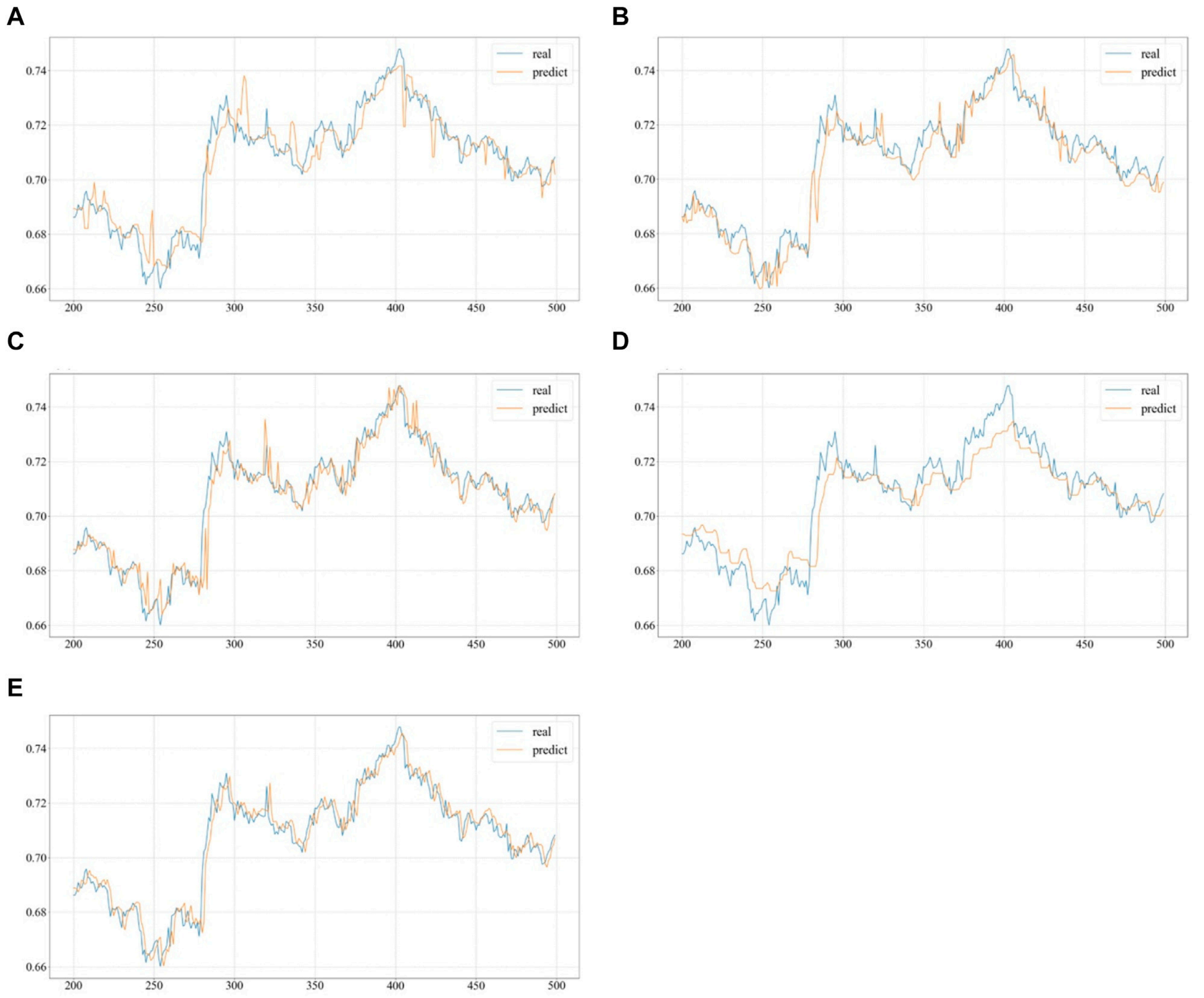


FIGURE 4
CNN–LSTM test results. **(A)** CNN-LSTM-base test results. **(B)** CNN-LSTM-ECA test results. **(C)** CNN-LSTM-SE test results. **(D)** CNN-LSTM-CBAM test results. **(E)** CNN-LSTM-TPA test results.

TABLE 4 Test results.

Model	RMSE	Reduce
CNN–LSTM–BASE	0.01254	0
CNN–LSTM–ECA	0.00526	58%
CNN–LSTM–SE	0.00036	97%
CNN–LSTM–CBAM	0.00023	98%
CNN–LSTM–TPA	0.00015	99%

$$L_{th2} = 0.5 * k * L_{20}(S). \quad (4)$$

In the field of tunnel vehicle operation data collection and processing, the integration of LiDAR-vision machines and edge processors can be utilized to sequentially accomplish the collection, recognition, and fusion of radar and video data. The steps are as follows: 1) The LiDAR-vision machine conducts real-time data collection of vehicles in the target area, acquiring both radar and video detection data. 2) The edge processor extracts radar and video detection data separately. It utilizes the YOLOv5 algorithm to extract vehicle type information from the video, and the 3DSSD radar target detection algorithm to extract vehicle position information. 3) The extracted target data undergo spatiotemporal synchronization. Through time registration, ineffective radar and video frames are eliminated. Spatial calibration is then applied to transform valid radar data into pixel space. 4) The region of interest (ROI) method is used to merge radar- and video-detected vehicle targets. Vehicle target information from both LiDAR and vision sources is fused based on detection distance, thereby achieving holographic perception of vehicles passing through highway tunnels.

However, the current operating environment in highway tunnels is harsh, with frequent malfunctions of operational data

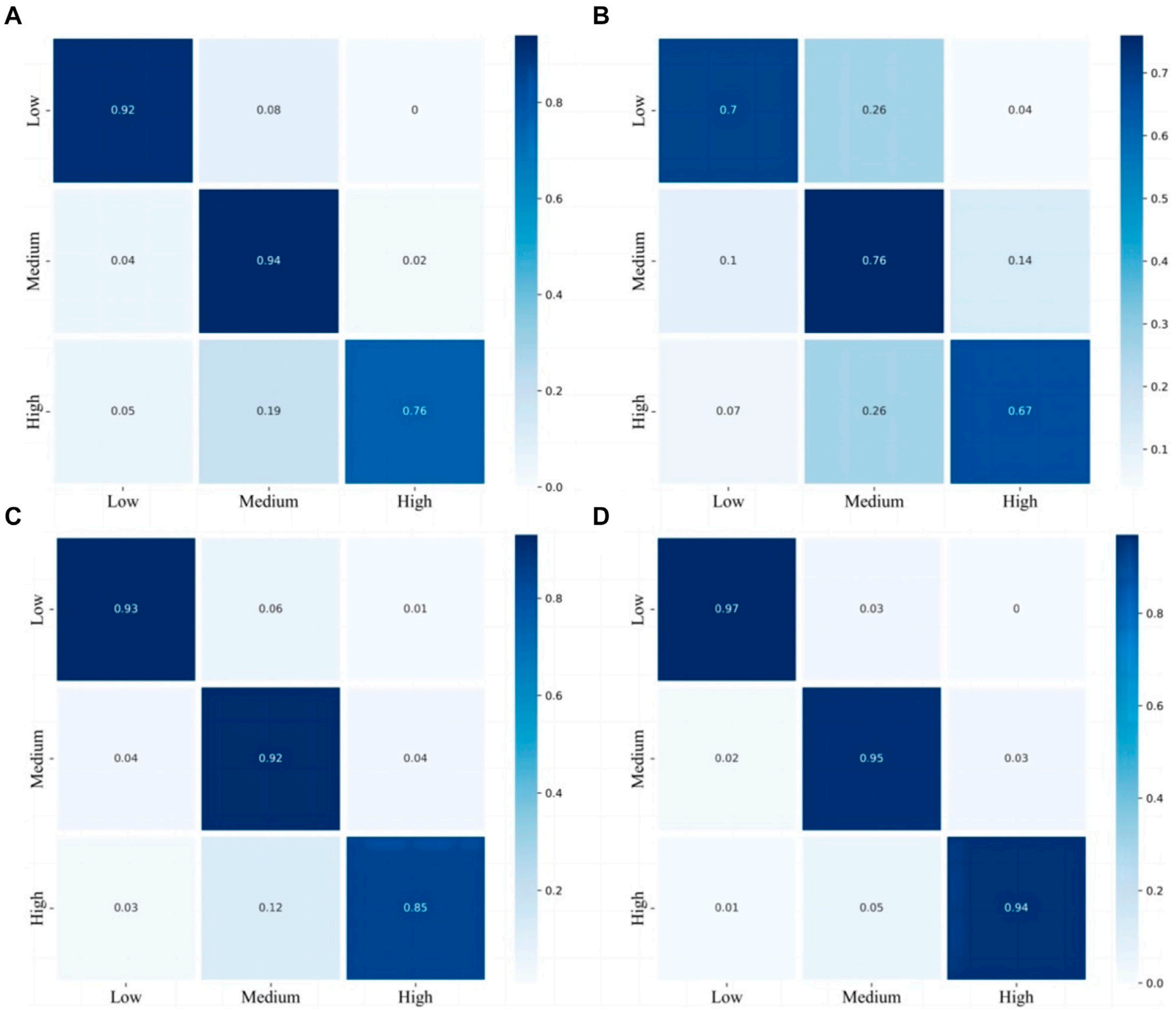


FIGURE 5
Confusion matrix. **(A)** CNN-LSTM-SE test results. **(B)** CNN-LSTM-ECA test results. **(C)** CNN-LSTM-CBAM test results. **(D)** CNN-LSTM-TPA test results.

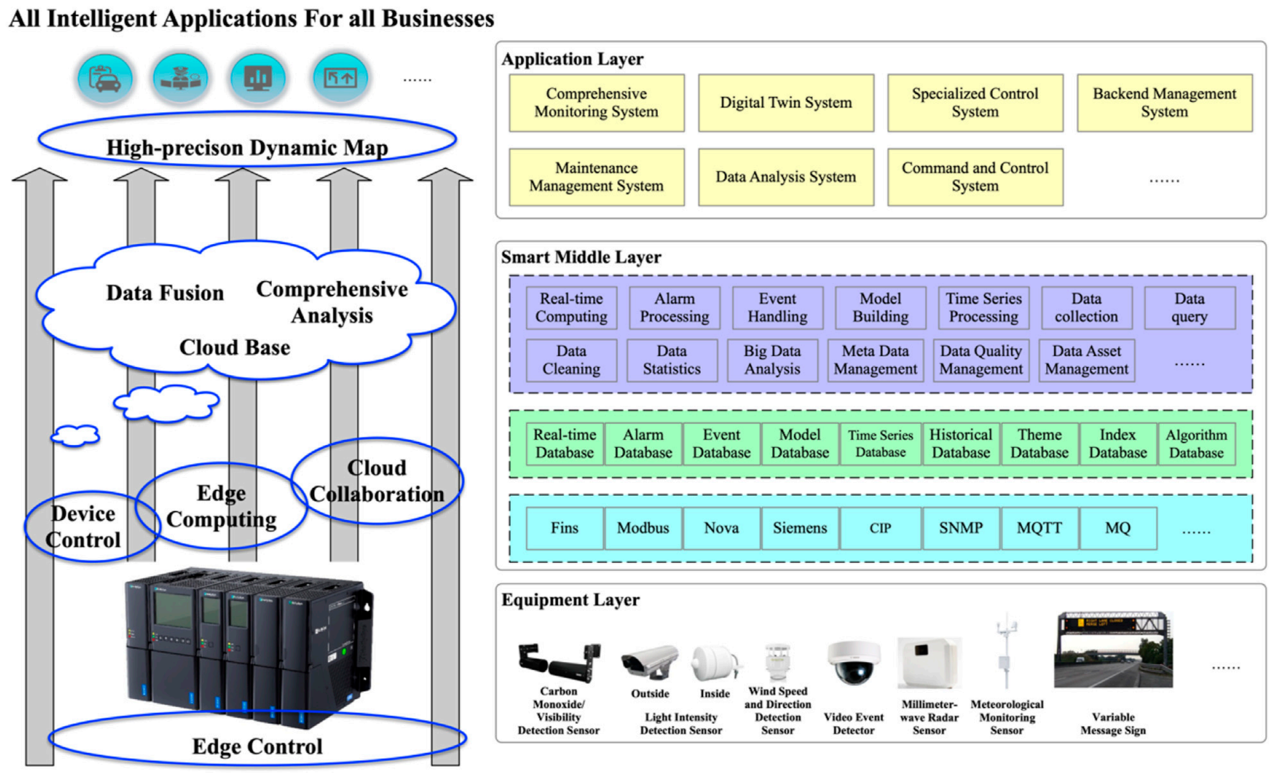


FIGURE 6
Schematic diagram of the overall system architecture.

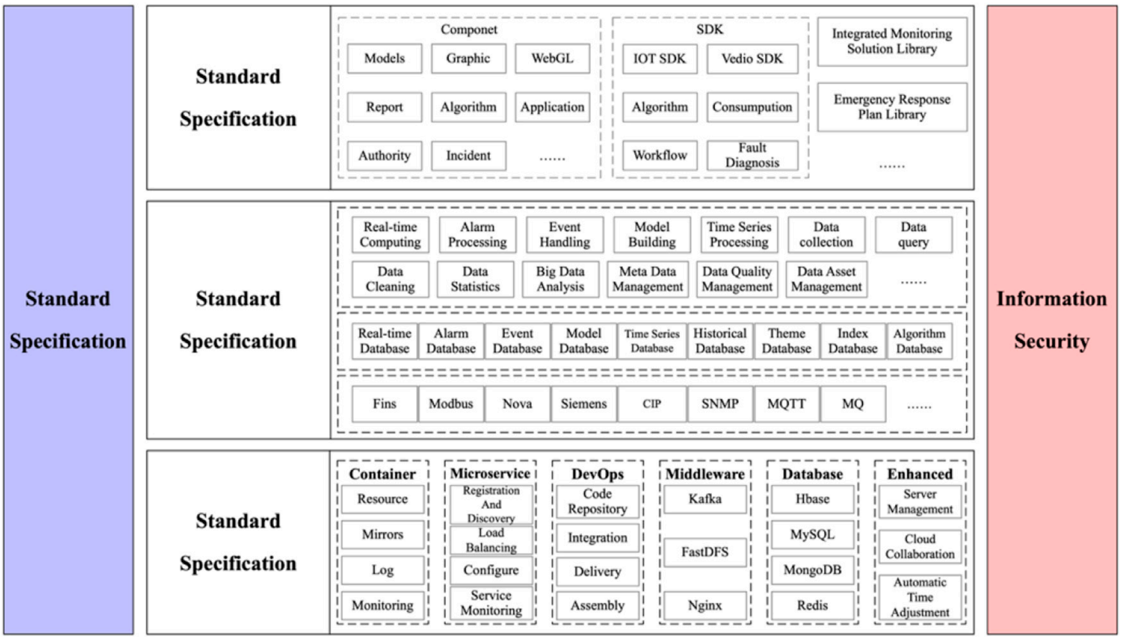


FIGURE 7
Schematic diagram of system technical architecture.

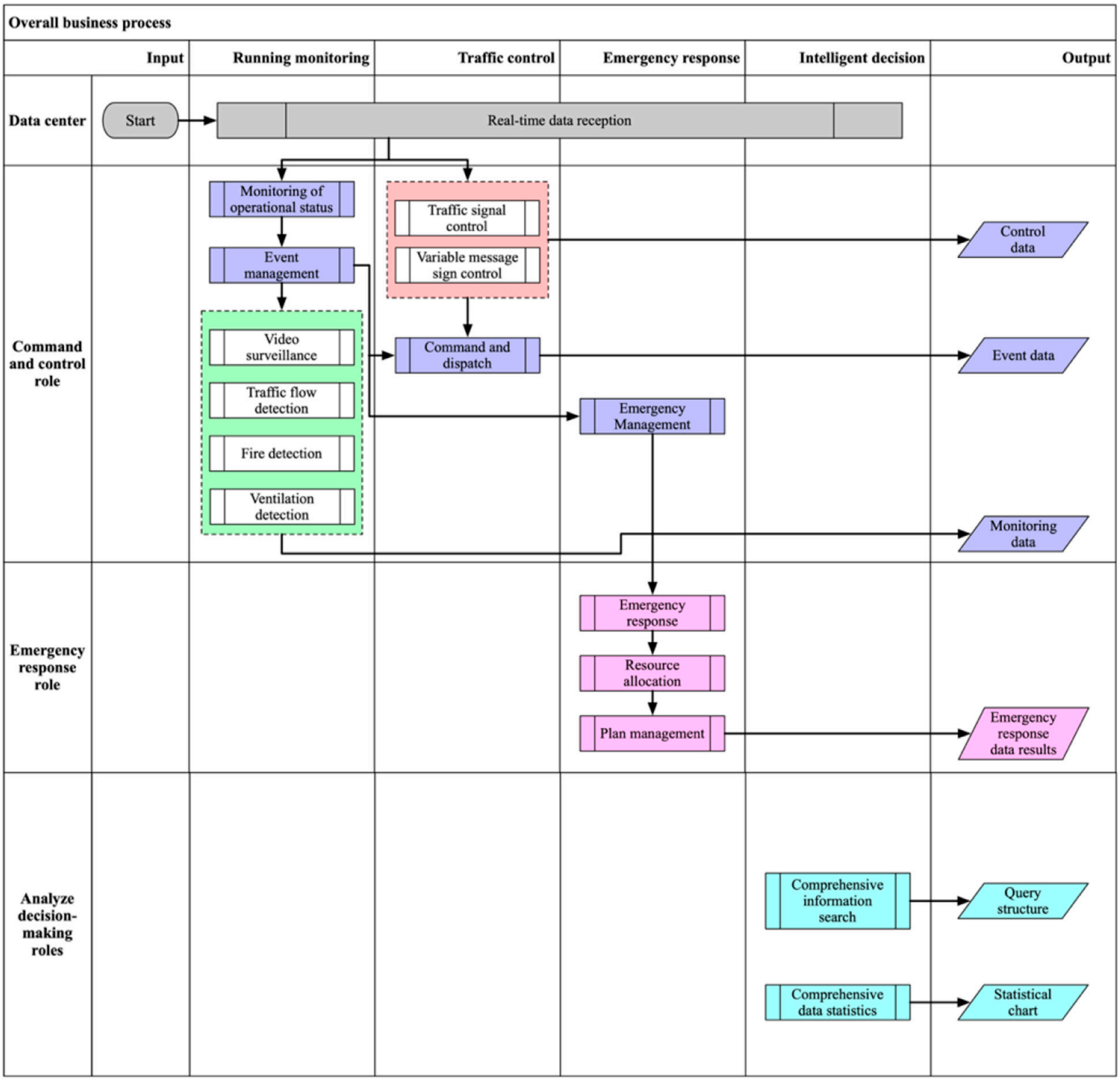


FIGURE 8
Schematic diagram of system business architecture.

detection equipment such as carbon monoxide/visibility detection sensors, wind speed/direction detection sensors, brightness detection sensors, and vehicle detection sensors, making long-term stable operation impossible. While video imaging can be used to detect traffic flow, visibility, and other data, it is still difficult to accurately predict the overall operating status of the tunnel, and operational management decisions lack effective data support. Due to most tunnel data detection sensors currently being integrated into the tunnel monitoring system via PLC controllers, the data are first converted from digital to analog form and then back to digital form before being transmitted to the monitoring system. If the PLC devices lack effective maintenance, the precision of the data will not meet the requirements for tunnel operation management [34, 35]. In response to these issues, this article proposes a method based on multi-sensor fusion to discriminate the operating state of highway tunnels.

3 Prediction method of highway tunnel operation status based on multimodal data fusion

3.1 Highway tunnel operation status prediction model based on multimodal data fusion

The operational status of highway tunnels encompasses various aspects such as traffic operation status, the adaptability of traffic

TABLE 5 Code of basic tunnel parameters.

Name	Abbreviation	Data types	Can be null	Notes
Tunnel name	TunName	Varchar (20)	False	
Tunnel beginning number	TunBegin	Numeric (7.3)	False	Unit: kilometer
Tunnel ending number	TunEnd	Numeric (7.3)	False	Unit: kilometer
Tunnel central number	TunCentre	Numeric (7.3)	False	Unit: kilometer
Classification code	ClaCode	Varchar (20)	False	
Length	Length	Numeric (6.2)	False	Unit: kilometer
Clear width	CleWidth	Numeric (6.2)	False	Unit: meter
Clear height	CleHeight	Numeric (6.2)	False	Unit: meter
Hole mode	HoMode	Varchar (20)	False	
Mode of the cross-section	SecMode	Varchar (20)	False	
Lining material	LinMaterial	Varchar (20)	False	
Mode of lighting conditions	LighMode	Varchar (20)	False	
Mode of ventilation	VenMode	Varchar (20)	False	
Mode of electromechanical facilities	FaciMode	Varchar (20)	False	
Completion date	ComDate	Datetime	False	
Design unit	DesUnit	Varchar (20)	False	
Construction unit	ConsUnit	Varchar (20)	False	
Supervision unit	SupUnit	Varchar (20)	False	
Management unit	ManUnit	Varchar (20)	False	
Maintenance unit	MainUnit	Varchar (20)	False	
Name of the sender	SendMan	Varchar (20)	Yes	
Date and time	SectTime	Datetime	Yes	
Vehicle detection sensor ID	VDID	Varchar (20)	False	
Collection time	RecTime	Datetime	False	
Collection cycle	RerPeriod	Smallint	Yes	
Upstream heavy vehicle flow	UupFluxB	Smallint	Yes	
Upstream light vehicle flow	UupFluxS	Smallint	Yes	
Upstream flow	UupFlux	Smallint	Yes	Total traffic volume of all lanes in the upstream direction
Downstream heavy vehicle flow	DwFluxB	Smallint	Yes	
Downstream light vehicle flow	DwFluxS	Smallint	Yes	
Downstream flow	DwFlux	Smallint	Yes	Total traffic volume of all lanes in the downstream direction
Upstream average speed	UpSpeed	Smallint	Yes	
Downstream average speed	DwSpeed	Smallint	Yes	
Upstream average occupancy rate	UpOccup	Numeric (5.2)	Yes	
Downstream average occupancy rate	DwOccdown	Numeric (5.2)	Yes	
Total number of lanes	LaneNum	Tinyint	Yes	Number of lanes detected by the equipment
Working status	Status	Tinyint	Yes	0- normal, 1- fault, and 2- unknown
Communication status	CommStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown

(Continued on following page)

TABLE 5 (*Continued*) Code of basic tunnel parameters.

Name	Abbreviation	Data types	Can be null	Notes
Carbon monoxide and visibility detection sensor ID	COVID	int	False	
Collection time	COVTime	Datetime	False	
Collection period	COVPeriod	Smallint	Yes	
Carbon monoxide concentration	COConct	Smallint	Yes	
Visibility	Visibility	Smallint	Yes	
Working status	WorkStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown
Communication status	CommStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown
Light intensity detection sensor ID	LOLIID	int	False	
Acquisition time	LOLITime	Datetime	False	
Acquisition period	LOLIPeriod	Smallint	Yes	
Outside brightness of the hole	LOLumi	Smallint	Yes	
Inside brightness of the hole	LILumi	Smallint	Yes	
Working status	WorkStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown
Communication status	CommStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown
Wind speed and direction detection sensor ID	WSID	int	False	
Collection time	WSTime	Datetime	False	
Collection cycle	WSPeriod	Smallint	Yes	
Wind direction	Direction	Tinyint	Yes	
Wind speed	Speed	Smallint	Yes	
Working status	WorkStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown
Communication status	CommStatus	Tinyint	Yes	0- normal, 1- fault, and 2- unknown

engineering and auxiliary facilities, the environmental adaptability of tunnel operations, and operational risk. These aspects can be quantitatively assessed through the status of tunnel infrastructure and data collected by sensors. The traffic operation status of tunnels can be determined using data from vehicle sensors. The adaptability of traffic engineering and auxiliary facilities can be assessed through the status data of tunnel electromechanical facilities. The environmental adaptability of tunnel operations can be evaluated using data collected by environmental sensors in the tunnel. Operational risks can be identified through event detection sensors. A comprehensive evaluation standard for the operational status of highway tunnels can be computed using multimodal data processing methods, which analyze the connections between various types of data.

This article combines the CNN–LSTM deep learning model with the self-attention mechanism to apply it to the judgment of tunnel operation status. The CNN–LSTM model is used to extract features from nonintrusive multimodal time series data, and the self-attention mechanism is used to integrate traffic flow, carbon monoxide, and visibility detection data, to effectively judge the tunnel operation service level by weighing the features of different modes [36]. The multimodal fusion architecture mainly includes four steps: preprocessing, feature extraction, feature fusion, and classification, as shown in Figure 2.

3.1.1 1D-CNN

The CNN model usually consists of three main components: the convolutional layer, pooling layer, and fully connected layer.

The role of the convolutional layer is to perform convolutional operation between the local region of the input data and the convolution kernel and slide the convolution kernel window to traverse the entire input data through local receptive fields. The convolution calculation equation is as follows:

$$\boldsymbol{x}_i^l = \boldsymbol{f}\left(\boldsymbol{w}_i^l * \boldsymbol{X}^{l-1} + \boldsymbol{b}_i^l\right). \tag{5}$$

In the equation, x_i^l represents the ith feature of the output value of layer l, w_i^l represents the weight matrix of the ith convolution kernel in layer l, * operator represents the convolution operation, X^{l-1} represents the output of layer l-1, b_i^l represents the bias term, and f represents the activation function of the output. CNN uses a nonlinear activation function to solve real-world nonlinear problems and chooses rectified linear unit (ReLU) as the activation function of the convolutional neural network.

The role of the pooling layer is to combine spatially, reducing the dimensionality of the feature map while maintaining the most important information. There are many types of it, and the maximum pooling is generally used, and its expression is

TABLE 6 Description of the digital twin system for highway tunnel operation.

Subsystem	Function name	Description
Comprehensive monitoring system	Tunnel daily management control	According to the monitoring center, the selected tunnels within its jurisdiction can realize the functions of tunnel electromechanical equipment status, detection information collection, single control, group control, event monitoring, and alarm information confirmation, and event and fault information entry in a three-dimensional and two-dimensional visual model
Digital twin system	Tunnel basic information digital twin	The digital twin presents the basic information of the tunnel
Digital twin system	Field electromechanical equipment digital twin	The digital twin presents and controls the electromechanical equipment outside the tunnel and related road sections
Digital twin system	Comprehensive environmental information digital twin	The digital twin presents the environmental monitoring information of the tunnel and related road sections
Digital twin system	Real-time traffic operation digital twin	The digital twin presents the real-time traffic flow and vehicle information of the tunnel and related road sections
Digital twin system	Traffic incident digital twin	The digital twin presents the tunnel event detection
Digital twin system	Emergency linkage digital twin	The digital twin presents the emergency linkage control plan of the tunnel
Specialized control system	Video inspection special item	The cameras of the selected road sections and tunnels are grouped into 16 video streams for broadcasting, and the situation inside the tunnel is inspected
Specialized control system	Tunnel lighting special item control	Remote control, manual control, intelligent control, and contingency control can be selected for the selected tunnels' lighting control
Specialized control system	Road guidance special item control	Graphically display the variable information identification settings of the tunnel's surrounding road network, display the current display content of each variable information sign, and support manual and contingency information release and single or group release per the contingency plan
Specialized control system	Electromechanical equipment linkage control	The linkage control plan can be customized based on the tunnel's actual needs, and the control modes are accident linkage control mode and daily linkage control mode
Command and control system	Linkage emergency plan management	The graphical interface realizes the linkage control plan for tunnel electromechanical equipment, and add/delete/modify/query functions are available
Command and control system	Emergency special plan management	The graphical interface realizes the emergency plan for tunnel events, and add/delete/modify/query functions are available
Command and control system	Operation log	The operation records of the current system users can be viewed
Maintenance management system	Maintenance task management	Tunnel electromechanical system maintenance task management (daily inspection, regular maintenance, and periodic maintenance task formulation) and tunnel maintenance plan formulation
Maintenance management system	Electromechanical equipment fault management	Manage faulty electromechanical equipment and fault repair tasks
Maintenance management system	Data management	Manage tunnel electromechanical system-related contracts and knowledge base
Maintenance management system	Operation log	The operation records of the current system users can be viewed
Data analysis system	Operation theme data statistical analysis	By collecting, summarizing, comparing, and analyzing operation-related data, statistical and analytical reports in predetermined or customizable formats can be generated
Data analysis system	Traffic theme data statistical analysis	By collecting, summarizing, comparing, and analyzing traffic-related data, statistical and analytical reports in predetermined or customizable formats can be generated
Data analysis system	Equipment data statistical analysis	By collecting, summarizing, comparing, and analyzing electromechanical equipment-related data, statistical and analytical reports in predetermined or customizable formats can be generated
Data analysis system	Dashboard- operation data display	Traffic theme, energy-saving theme, environmental theme, and equipment status data display
Data analysis system	Dashboard- electromechanical equipment automatic inspection	Inspect the working status of tunnel electromechanical equipment online and automatically discover abnormal devices

(Continued on following page)

TABLE 6 (*Continued*) Description of the digital twin system for highway tunnel operation.

Subsystem	Function name	Description
Data analysis system	Dashboard- emergency command	Event detection, alarm confirmation, video call, and contingency plan selection and demonstration
Data analysis system	Operation log	The operation records of the current system users can be viewed
Backend management system	Role information management	The platform's organization structure management, personnel management, full selection management, and user management are available
Backend management system	Basic data management	Manage electromechanical equipment, contract unit, equipment manufacturer, emergency facilities, and external units management
Backend management system	Platform log management	The operation records of the current system users can be viewed

$$y_i^{l+1}(j) = \max x_i^j(k) k \in D_j. \tag{6}$$

This sentence describes the max pooling operation in a CNN model. $y_i^{l+1}(j)$ represents an element in the jth pooled feature map of the (l+1)th layer after pooling. D_j is the jth pooling region, and $x_i^j(k)$ represents an element of the lth layer's ith feature map within the pooling kernel.

3.1.2 LSTM

LSTM is an upgraded variant of RNN that adds gate structures internally, including input gates, forget gates, and output gates, which can adjust the values of input and hidden layers. The calculation process is as follows:

$$\begin{cases} f_t = \sigma\left(W_f\left[h_{t-1}, x_t + b_f\right]\right) \\ i_t = \sigma\left(W_i\left[h_{t-1}, x_t + b_i\right]\right) \\ \tilde{C}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \\ C_t = f_t C_{t-1} + i_t \tilde{C}_t \\ o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ h_t = \sigma_t \tanh\left(C_t\right) \end{cases}. \tag{7}$$

In the equation, f_t, i_t, and o_t are the computation functions of the forget gate, input gate, and output gate vectors, respectively. $\tilde{C}_t$ represents the candidate state information. σ is the sigmoid function. W_f, W_i, W_c, and W_o are the corresponding weight matrices, and b_f, b_i, b_c, and b_o are the corresponding bias vectors. x_t is the input vector at time t, C_{t-1} is the stored cell information from the previous time step, and h_t is the LSTM output vector.

3.1.3 Attention mechanism

By using the attention mechanism, efficient allocation of information processing resources can be achieved. Due to the difference in the importance of features in short subsequences of long time series, significant features often contain more information and have a greater impact on the trend of actual demand. If CNN is given the ability to focus more on high-importance features, it can better extract short-term patterns and optimize LSTM input information [37]. Therefore, this paper uses the attention mechanism to extract significant features of short sequences.

The attention mechanism can be categorized into hard attention and soft attention. Hard attention selects the ROI as input and is effective in focusing on the target object by removing meaningless background data in image research. However, the direct restriction of input content processing method used in hard attention is not entirely suitable for time series prediction. Even if there are differences in the importance of input sequences, since each input subsequence contains certain information at different positions in the sequence, it cannot be identified and removed. Additionally, hard attention requires reinforcement learning optimization, which makes training difficult and less universal. In contrast, soft attention uses weights trained by neural networks to globally weight input features in space or channel, achieving the goal of focusing on specific spatial regions or channels. Moreover, this method is differentiable in backpropagation, allowing end-to-end learning and direct learning of attention networks. Based on these principles, this paper introduces soft attention into one-dimensional CNN, weighting all input features one by one, focusing on specific spatial regions and channels to achieve the significant and fine-grained feature extraction of time series.

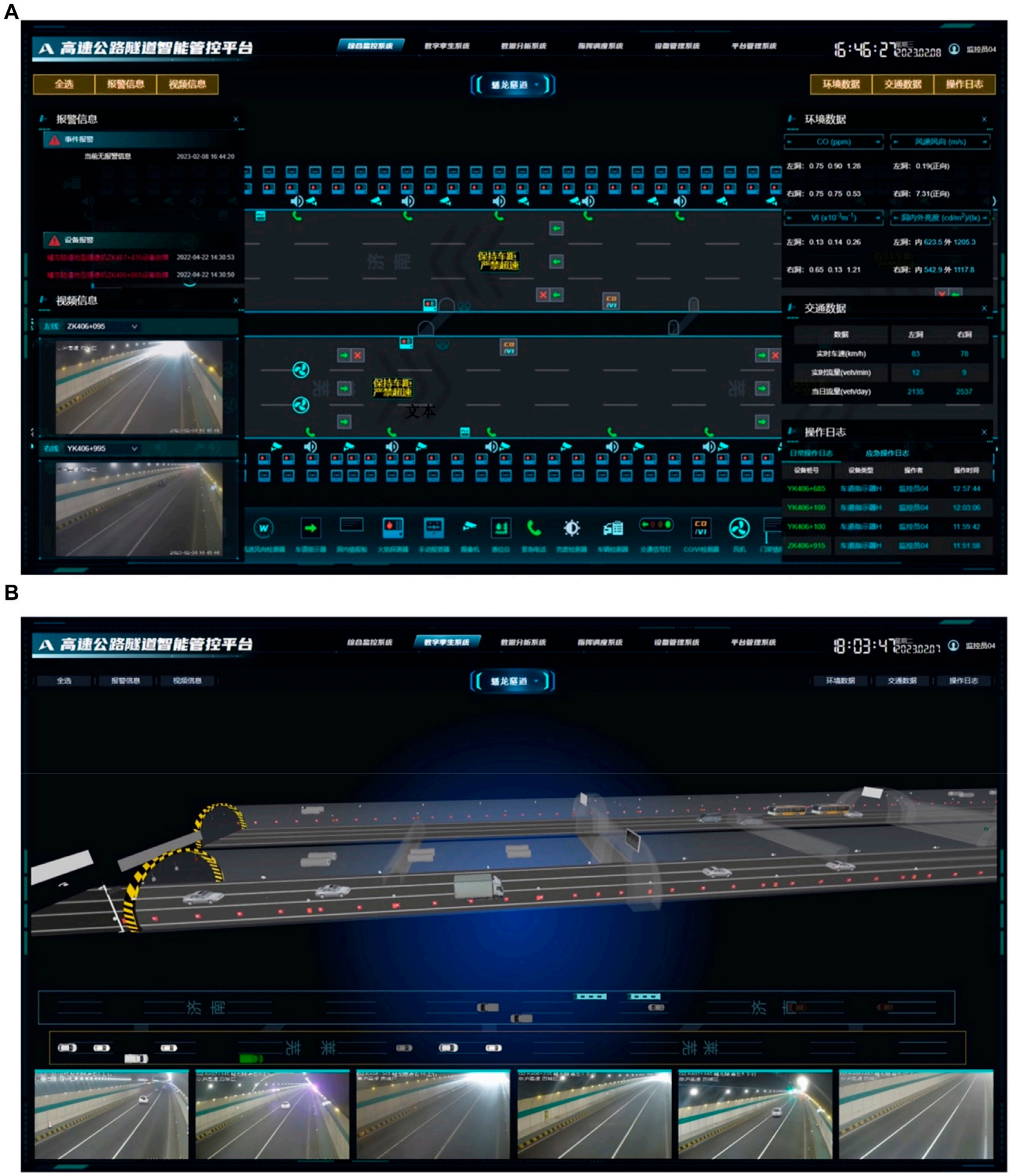
3.1.3.1 SE attention mechanism

The purpose of the SE (squeeze-and-excitation) module is to apply a weight matrix from the channel domain perspective, assigning different weights to various positions in an image, thereby extracting more significant feature information.

To obtain channel-wise attention, the feature map is first globally average pooled based on its width and height, reducing spatial dimensions to 1 × 1. Then, two fully connected layers and nonlinear activation functions are used to establish connections between channels. The SE module first performs a "squeeze" operation on the convolutional feature map to obtain global channel-wise features and then performs an "excitation" operation to learn the relationships between channels and obtain weights for each channel. Finally, the original feature map is multiplied by the channel-wise weights to obtain the final feature. Essentially, the SE module performs attention operation on the channel dimension, allowing the model to focus more on the most informative channel features while suppressing those that are not important.

3.1.3.2 ECA attention mechanism

The SE attention mechanism first compresses the input feature map along the channel dimension, but this compression can have a negative impact on learning dependencies between channels. Based on this idea, the ECA attention mechanism avoids dimensionality reduction and efficiently implements local cross-channel interactions using a 1D convolution to extract inter-channel dependencies. The specific steps are as follows:



报警信息

当前无报警信息 2023-02-08 16:44:20

2022-04-22 14:30:53

2022-04-22 14:30:50

环境数据

CO (ppm)

左洞: 0.75 0.90 1.28

右洞: 0.75 0.75 0.53

风速风向 (m/s)

左洞: 0.19(正向)

右洞: 7.31(正向)

VI ($x10^{-3}m^{-1}$)

左洞: 0.13 0.14 0.26

右洞: 0.65 0.13 1.21

洞内外亮度 (cd/m^2)/(lx)

左洞: 内 623.5 外 1205.3

右洞: 内 542.9 外 1117.8

交通数据

数据	左洞	右洞
实时车速(km/h)	83	78
实时流量(veh/min)	12	9
当日流量(veh/day)	2135	2537

操作日志

日常操作日志　应急操作日志

设备桩号	设备类型	操作者	操作时间
YK406+685	车道指示器H	监控员04	12:57:44
YK406+100	车道指示器H	监控员04	12:03:06
YK406+100	车道指示器H	监控员04	11:59:42
ZK406+915	车道指示器H	监控员04	11:51:58

FIGURE 9
Digital twin system for highway tunnel operation. **(A)** Integrated monitoring system function page. **(B)** Traffic operation digital twin page.

Step 1: Global average pooling is performed on the input feature map.

Step 2: A 1D convolution operation is performed with a kernel size of k, and the sigmoid activation function is applied to obtain the weight w for each channel, as shown in the following equation:

$$\omega = \sigma\left(C1D_k\left(y\right)\right). \tag{8}$$

Step 3: The weights are multiplied with the corresponding elements of the original input feature map to obtain the final output feature map. The idea and operation of the ECA attention mechanism are extremely simple and have minimal impact on network processing speed. However, ECA attention only uses channel attention, and its accuracy still needs to be verified for specific application scenarios.

3.1.3.3 Convolutional block attention module

The convolutional block attention module (CBAM) combines the two-dimensional attention mechanism of feature channel and feature space, and the structure diagram is shown in Figure 3.

The CBAM, like SE-Net, automatically learns the importance of each feature channel. Additionally, it learns the importance of each feature space in a similar manner. By utilizing the importance levels obtained, the CBAM enhances relevant features and suppresses those less important for the current task.

The CBAM extracts channel attention in a manner largely similar to SE-Net, as demonstrated in the code for channel attention. Building upon the foundation of SE-Net, the CBAM introduces an additional feature extraction method using max pooling, while the remaining steps are identical. The features extracted from channel attention serve as inputs for the spatial attention module.

In the CBAM, the method for extracting feature space attention involves processing the feature maps through channel attention to prioritize channels based on their importance. These feature maps are then fed into the spatial attention module. Similar to the channel attention module, spatial attention involves processing the channels through both maximum and average pooling. The results of these two processes are concatenated, followed by a convolutional operation to reduce them into a 1WH feature map, representing spatial weights. These weights are then applied to the input features through a point-wise multiplication, thereby implementing the spatial attention mechanism.

3.1.3.4 Temporal pattern attention mechanism

Temporal pattern attention (TPA) is used for multivariate time series forecasting. First, a large number of time series are fed into LSTM to obtain a hidden state matrix H. For each row (*i*th row) of the hidden state matrix H, k CNN filters are used to extract features, resulting in an $n*k$-dimensional matrix H_C.

$$\boldsymbol{H}_{i,j}^{C} = \sum\nolimits_{l=1}^{w} \boldsymbol{H}_{i}, (\boldsymbol{t} - \boldsymbol{w} - \boldsymbol{1} + \boldsymbol{l}) * \boldsymbol{C}_{j,T-w+l}. \tag{9}$$

For h_t to be predicted, it is interacted with each row of the H_C matrix to produce a weight a_i for each row. This weight represents the strength of the effect of each row of the H_C matrix on h_t to be predicted, i.e., the strength of the influence of each time series on h_t.

$$\boldsymbol{f}\left(\boldsymbol{H}_{i}^{C}, \boldsymbol{h}_{t}\right) = \left(\boldsymbol{H}_{i}^{C}\right)^{T} \boldsymbol{W}_{a} \boldsymbol{h}_{t}, \tag{10}$$

$$\boldsymbol{a}_{i} = \boldsymbol{sigmoid}\left(\boldsymbol{f}\left(\boldsymbol{H}_{i}^{C}, \boldsymbol{h}_{t}\right)\right). \tag{11}$$

Each row is weighted and summed to obtain v_t, which represents the combined effect of all rows on h_t, i.e., the effect of time, i.e., time attention.

$$\boldsymbol{v}_{t} = \sum\nolimits_{i=1}^{n} \boldsymbol{a}_{i} \boldsymbol{H}_{i}^{C}. \tag{12}$$

When predicting h_t, we add the influence of all time series on h_t to the original equation, namely,

$$\boldsymbol{h}_{t}^{,} = \boldsymbol{W}_{h} \boldsymbol{h}_{t} + \boldsymbol{W}_{v} \boldsymbol{v}_{t}, \tag{13}$$

$$\boldsymbol{y}_{t-1+\Delta} = \boldsymbol{W}_{h^{,}} \boldsymbol{h}_{t^{,}}. \tag{14}$$

The first step is to synchronize the dynamic traffic detection data and environmental detection data. Abnormal data in the detection data are identified and replaced or removed intelligently. To reduce the data differences between different monitoring points, all data are normalized. Then, the sliding window method is used to divide each feature of each mode into time windows with a fixed window size and overlap. A new training dataset is composed of the generated time windows, with each label corresponding to the original dataset.

Next, the new training datasets for each mode are input into the 1D-CNN and LSTM framework to extract features. Segment time windows from the training dataset are first fed into the 1D-CNN to automatically learn features. Since the time window is a time series, a one-dimensional convolution layer is used. The feature extraction framework consists of three one-dimensional convolution layers, three max pooling layers, and two LSTM layers, with detailed parameter settings shown in Table 3. The convolution layer uses a sliding filter to extract effective features. The activation function of the convolution layer is chosen as the exponential linear unit (ELU), which can accelerate convergence and improve the robustness of the model. After each convolution layer, a max pooling layer is used to reduce the amount of data to half the original size. A dropout layer is used after the pooling layer to avoid overfitting. In each training epoch, a random subset of the neurons in the dropout layer is selected and not allowed to participate in weight optimization. After three layers of convolution and pooling, the input data are transformed into a high-dimensional feature map. Since the feature map is extracted from the time window, and the convolution and pooling operations do not change their time sequence, the feature map is directly input into two LSTM layers. The LSTM network handles time series through gate mechanisms, including forget gates, input gates, and output gates. They can control the discard or addition of information to achieve forgetting and memory. The LSTM network converts the feature map into the corresponding hidden state.

During the fusion step, the hidden states generated from the detection data are integrated to create a new feature map. This feature map contains hidden states and is denoted as H:

$$\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n). \tag{15}$$

Due to the varying degrees of impact of different hidden states on tunnel operation monitoring, this paper introduces a self-attention mechanism to measure all hidden states. These hidden states are aggregated into a vector s through an attention layer, which is calculated using the following equation:

$$\boldsymbol{u}_t = \textbf{tanh}(\boldsymbol{w}\boldsymbol{h}_t + \boldsymbol{b}), \tag{16}$$

$$\boldsymbol{a}_t = \frac{\textbf{exp}\left(\boldsymbol{u}_t^{T}\boldsymbol{u}\right)}{\sum_{t=1}^{n}\textbf{exp}\left(\boldsymbol{u}_t^{T}\boldsymbol{u}\right)}, \tag{17}$$

$$\boldsymbol{s} = \sum\nolimits_{i=1}^{n} \boldsymbol{a}_t \boldsymbol{h}_t. \tag{18}$$

The hidden state h_t is first input into a fully connected layer with a tanh activation function to obtain the hidden representation u_r as h_t. The transpose of the output values is multiplied by a trainable parameter vector to obtain the attention alignment coefficients. Then, the alignment coefficients are normalized using the softmax function to obtain the summation weights a_t. Next, the vector representation s is computed as the weighted sum of the hidden states. In the final step of the decision-making process, the vector representation s can be input as a feature vector into a softmax classifier for judgment. Here, w is the weight matrix and b is the bias vector of the fully connected layer in the attention layer, with a dimension of d_a. The parameter vector u represents the context information and also has a dimension of d_a. The value of d_a is an important hyperparameter of this model; therefore, to balance model performance and computational complexity, the optimal dimension of d_a is set to 64. During the training process, the weight matrix w, bias vector b, and parameter vector u are randomly initialized.

3.2 Example analysis

This article focuses on predicting the air quality level in tunnel operation using environmental detection data (carbon monoxide, visibility) and traffic detection data (vehicle flow) as the research object. The dataset consists of 10,500 sets of data collected automatically every 5 min from carbon monoxide/visibility sensors and integrated detection sensors, at the same location and time in the tunnel. Among them, 10,000 sets of data are used as training samples, and the remaining 500 sets are used as test samples. The sample data are combined into a feature vector using a CNN–LSTM–attention model, and this feature vector is set as the air quality level of tunnel operation to compare the predicted and actual states to test the accuracy of the proposed multimodal fusion algorithm for evaluating tunnel operation status.

3.2.1 Prediction process

The feature extraction framework for traffic detection data and environmental detection data mainly includes three steps: preprocessing, feature extraction, feature fusion, and classification. First, all data are normalized using min–max normalization. After preprocessing (missing and abnormal values), three types of environmental features (carbon monoxide, visibility, and vehicle flow) and traffic features are extracted. Finally, all traffic and environmental features are combined into a feature vector, which is then input into the classifier. A CNN–LSTM–attention network is used as the classifier for handcrafted traffic and environmental features.

The handcrafted multimodal data fusion method combines features from traffic detection data and environmental detection data into a feature vector, which is set as the air quality level of tunnel operation. Based on the corresponding historical air quality level of tunnel operation for traffic and environmental detection data at the same location and time in the tunnel, this feature vector is defined as a value between 0 and 1, where 0–0.4 is low, 0.4–0.8 is medium, and 0.8–1.0 is high. This vector is input into the CNN–LSTM–attention network, and the predicted air quality level of tunnel operation is compared with the actual state.

3.2.2 Prediction results

The CNN–LSTM–attention network effectively fuses data from different modalities by allocating different weights to different features through the self-attention mechanism. The prediction results of different attention mechanisms, including CNN–LSTM–BASE, CNN–LSTM–ECA, CNN–LSTM–SE, CNN–LSTM–CBAM, and CNN–LSTM–TPA, without adding the self-attention mechanism are compared, and the samples between 200 and 500 are selected. The accuracy of the evaluation algorithm under different attention mechanisms is different. The prediction results of each model are shown in Figure 4.

3.2.3 Evaluation metrics

The root mean square error (RMSE) is used as the measure of accuracy, which is the square root of the sum of the squared differences between the predicted values and the actual values, divided by the number of observations m.

$$\boldsymbol{RMSE}(\boldsymbol{X}, \boldsymbol{h}) = \sqrt{\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{i=1}^{m}\left(\boldsymbol{h}\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{y}^{(i)}\right)^2}. \tag{19}$$

3.2.4 Evaluation results

The evaluation results of each model on the test set are shown in Table 4, and the confusion matrix is shown in Figure 5.

Based on the test results shown in Figure 5, the predictions of the CNN–LSTM–attention model for tunnel operation air quality levels are very close to the true values. According to the RMSE test results in Table 4, the CNN–LSTM–ECA, CNN–LSTM–SE, CNN–LSTM–CBAM, and CNN–LSTM–TPA models reduced RMSE by 58%, 97%, 98%, and 99%, respectively, compared to the CNN–LSTM–BASE model without the self-attention mechanism.

The confusion matrix shown in Figure 5 indicates that the CNN–LSTM–SE model achieved prediction accuracy rates of 0.76, 0.93, and 0.92 for high, medium, and low levels of tunnel operation air quality, respectively. This suggests that the SE attention mechanism can effectively predict the low and medium levels of tunnel operation air quality. The CNN–LSTM–ECA model achieved prediction accuracy rates of 0.76, 0.70, and 0.63 for high, medium, and low levels of tunnel operation air quality, respectively, indicating that the ECA attention mechanism had moderate performance in predicting the high, medium, and low levels of tunnel operation air quality. The CNN–LSTM–CBAM model achieved prediction accuracy rates of 0.85, 0.92, and 0.93 for high, medium, and low levels of tunnel operation air quality, respectively, indicating that the CBAM attention mechanism can effectively predict the low and medium levels of tunnel operation air quality. The CNN–LSTM–TPA model achieved prediction accuracy rates of 0.94, 0.95, and 0.97 for high, medium, and low levels of tunnel operation air quality, respectively, indicating that the TPA attention mechanism can effectively predict the high, medium, and low levels of tunnel operation air quality.

From the test results and confusion matrix, it can be concluded that the CNN–LSTM–attention model has high prediction accuracy for tunnel operation air quality levels, with an average NMSE of 0.0015 and an average reduction of 70%. The multimodal fusion algorithm using the TPA attention mechanism achieved the best analysis and prediction performance.

4 Design and development of the digital twin system for highway tunnel operation

4.1 System architecture design

4.1.1 Overall structure

The digital twin system for highway tunnel operation is based on the current situation of the highway tunnel operation management system and monitoring system. It fully utilizes existing resources and constructs a highway tunnel operation data system with data homogeneity and business collaboration. The system is not only suitable for the current distributed architecture model of regional controllers and services but also adaptable to the future smart highway's edge-cloud architecture [38–40], as shown in Figure 6. On the basis of ensuring the unity of data and business, the system can be efficiently iterated and updated to support the practical implementation of various innovative services for the future smart highway.

4.1.2 Technical architecture

The digital twin system for highway tunnels adopts a middleware architecture for design and development, which extracts reusable capabilities from the business, data, technology, algorithms, and other aspects of highway tunnel operations management to form a middleware platform, as shown in Figure 7.

(1) Basic backend

The basic backend fully considers various equipment interfaces and communication methods. By incorporating Internet of Things (IoT) access modules with built-in multi-brand and multi-type device communication methods and protocols (such as TCP, WebSocket, UDP, and HTTP), it ensures reliable and stable communication and fast integration of tunnel electromechanical equipment. After unifying coding standards, it forms the highway tunnel basic database, business database, theme database, and shared database.

(2) Capability middleware

Business middleware: Precious business capabilities are precipitated into the business middleware to achieve business capability reuse and linkage and coordination between various business modules, ensuring stable and efficient critical business links and enhancing business innovation efficiency.

Data middleware: Highway operating data are uniformly managed to provide complete and accurate data services for various business applications, including data storage, processing, and management.

Technology middleware: Common facilities, development technology components, and services are integrated and packaged to provide simple, consistent, and easy-to-use basic infrastructure capability interfaces, which help the rapid development of upper-layer services.

(3) Application frontend

The application frontend is built around the core business of tunnel operation management and includes various function systems such as the tunnel basic information digital twin, field electromechanical equipment digital twin, comprehensive environmental information digital twin, real-time traffic operation digital twin, abnormal traffic event digital twin, daily operation management digital twin, and emergency linkage control digital twin, achieving the digitalization, three-dimensionalization, and precision monitoring and management of tunnel operation.

4.1.3 Business architecture

In daily management, the comprehensive monitoring system is the business core, which collects real-time detection data of highway tunnel, monitors the operation status, and completes daily monitoring management of traffic control, ventilation, lighting, etc. In abnormal events, it realizes zoning and intelligent linkage control based on the location of the event, emergency plan, and precise implementation of special plans, as shown in Figure 8.

4.2 Digital twin model

The digital twin system for highway tunnels divides the multidimensional data of highway tunnel operation and management into basic parameter data, electromechanical facility operation and functional data, civil engineering structure facility condition data, event data, and maintenance inspection data. According to the actual needs of the information project, a digital twin model can be established for the relevant data. For the convenience of data interconnection, the data of the existing highway tunnel can be converted and coded according to the coding format, and the data of the new highway tunnel can be coded according to the coding format requirements. The data categories of the highway tunnel digital twin model should include tunnel basic information, electromechanical facility status and operation function data, event data, and maintenance inspection data. The operation data should be coded according to a unified standard, with complete and accurate parameter information and following the "one source, one number" principle to avoid duplicate collection. Data should be managed and classified in a centralized manner, and the application types can include equipment and facility operation monitoring, traffic safety control, abnormal event handling, and public travel services. The digital twin system for highway tunnels should achieve information interconnection, integration, sharing, and exchange. The code of highway tunnel information model is shown in Table 5.

In this article, 3D visualization modeling is conducted for the basic information, mechanical and electrical facilities, and operational environment detection data of the highway tunnel. The following are partial 3D visualization model prototype diagrams for some mechanical and electrical facilities.

4.3 System function research and development

The digital twin system for highway tunnels includes seven functional subsystems: comprehensive monitoring, digital twin, specialized control, maintenance management, command and dispatch, data analysis, and backend management. The subsystems share data and coordinate their operations to achieve unified control over 14 types of electromechanical equipment, such as ventilation, lighting, traffic control, and guidance. The system enables fine-tuned, modular, and intelligent control over ventilation and lighting dimming, as well as standardized and integrated classification control over the entire route guidance system. It also includes intelligent and linked control based on event types and rapid, standardized, and traceable command and dispatch based on eight types of frequently occurring event plans, as well as one-stop configuration management of basic data information.

The digital twin system for highway tunnel operations is a tangible representation of sensor data. Taking the precise perception of traffic volume data (such as cross-sectional traffic flow, regional traffic flow, vehicle violation information, and traffic event information) as an example, a holographic, visual, and digital model of vehicles passing through tunnels is established using vehicle positioning and trajectory fusion technologies. This model allows for the identification of vehicle violations such as wrong-way driving and illegal lane changes based on high-precision trajectory data and determines speeding or slow-moving vehicles based on their speed. Additionally, traffic event types are detected and classified using feature matching and deep learning techniques.

The multimodal information fusion algorithm plays a crucial role, especially when one or more sensors fail or provide erroneous detection data. By utilizing the spatiotemporal correlation between data, the algorithm calibrates and supplements problematic data, thereby deriving a comprehensive operational status evaluation indicator.

4.4 Engineering application verification

Multiple sensor fusion perception, multimodal data fusion, and digital twinning technologies have been applied to the highway tunnel operation control system. The system has been successfully implemented in more than 2,000 km of long tunnels in nine provinces, achieving the standardization of basic data, visualization of daily management, and process-based emergency control of tunnel operation management. The system collects highway tunnel operation monitoring data through multiple sensors and applies a multimodal information fusion method based on CNN–LSTM–attention to predict the highway tunnel operation status. The system supports the calculation of tunnel ventilation and lighting requirements under various operating conditions and improves the reliability and accuracy of tunnel operation intelligent control under normal conditions and can enable tunnel ventilation and luminaires to autonomously adjust based on external environmental changes. This adaptive regulation significantly reduces the energy consumption costs associated with tunnel operations. In tunnel lighting, lamps are the primary consumers of electrical energy. In practical applications, the system employs a smart lighting control method based on multi-parameter control. This approach effectively enhances the overall visual environment of the tunnel, reducing the adverse effects of tunnel black hole and white hole phenomena on driving safety. The lighting fixtures and other electromechanical facilities adaptively adjust according to external environmental changes, significantly reducing the energy consumption costs of tunnel operations. As a result, there is an approximate 20% reduction in the energy consumption costs of tunnel operation. By accurately monitoring tunnel environmental data and setting up fan interlocking control programs, it is possible to achieve a 100% qualification rate for air quality during the regular operation of the tunnel. In abnormal conditions, by proactively defining the control scope for tunnel emergencies, traffic management and electromechanical equipment interlock control plans are automatically generated based on real-time monitoring data (location and type of the event) for early warning. Once the monitor confirms the situation, they can simply click a confirmation button within the system to deploy the prearranged plan with a single click. This significantly enhances the capability for traffic accident prevention and control, as well as the emergency response to sudden incidents, ensuring that the emergency response time for exceptional events is less than 2 min. The number of tunnel traffic accidents has been reduced by over 25% for two consecutive years, effectively guaranteeing the “safe, smooth, and orderly” operation of the highway tunnel. Description of the digital twin system for highway tunnel operation is given in Table 6, and function pages of the system are shown in Figure 9.

5 Conclusion

This article tackles the challenge of frequent failures in operational data detection equipment within highway tunnels, including sensors for carbon monoxide/visibility, wind speed and direction, brightness, and vehicle detection. The harsh internal environment of tunnels makes accurate prediction of their operational state difficult, resulting in a lack of effective data support for management. To address these issues, this article advocates for the adoption of multi-sensor fusion perception and digital twin technology in the information infrastructure of highway tunnels. By creating a unified digital twin information model tailored to the tunnel’s operational characteristics, and applying a multimodal information fusion method based on CNN–LSTM–attention, the accuracy of highway tunnel operational status assessments has been markedly improved. This approach significantly enhances the stability and reliability of target recognition, reduces the likelihood of target omission, and, through data-driven methods, greatly improves the efficiency of tunnel ventilation and lighting control. The developed digital twin system for highway tunnels addresses centralized management, linkage control, data sharing, and business coordination challenges. Practical engineering results demonstrate that the system has bolstered tunnel traffic safety, reduced management costs, and improved the comfort of tunnel passage, thereby ensuring the “safety, smoothness, and orderliness” of highway tunnel operations.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

XY: conceptualization, methodology, and writing–original draft. SY: investigation and writing–original draft. JW: resources and writing–review and editing. HC: funding acquisition, resources, and writing–review and editing. YH: software and writing–review and editing. ZL: data curation and writing–review and editing. LF: funding acquisition and writing–review and editing.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Key Technology Projects in the transportation industry of China for 2021 (No. 2021-MS6-145) and the Key Research and Development Program of Guangxi, China (Grant No. AB21196034).

Acknowledgments

The authors would like to thank the support of their colleagues in the Research and Development Center of Transport Industry of Self-driving Technology.

Conflict of interest

Authors XY, SY, ZL, and LF were employed by China Merchants Chongqing Communications Research and Design Institute Co., Ltd. Author JW was employed by Gansu Provincial Highway Aviation Tourism Investment Group Co., Ltd. Author HC was employed by Shandong Hi-speed Company Limited. Author YH was employed by Guangxi Computing Center Co., Ltd.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2024.1335494/full#supplementary-material

References

1. China Journal of Highway And Transport Editorial Department. Review of academic research on China's traffic tunnel engineering, 2022. *China J Highw Transport* (2022) 35(4):1–14. doi:10.19721/j.cnki.1001-7372.2022.04.001

2. Tu Y, Wang S, Wang X, Shi L, Li M, Zhou G, et al. F L. Discussion on Quality Upgrading of Highway Tunnels. *Tunnel Construction* (2019) 39(S1):1–9. doi:10.3973/j.issn.2096-4498.2019.S1.001

3. Han Z, Guan F. Study on collaborative management of intelligent highway tunnels. *Tunnel Construction* (2018) 38(04):533–7.

4. Amundsen FH, Ranes G. Studies on traffic accidents in Norwegian road tunnels. *Tunnelling Underground Space Tech* (2000) 15(1):3–11. doi:10.1016/S0886-7798(00)00024-9

5. Geva V, Shinar D, Blum Y. In-vehicle information systems to improve traffic safety in road tunnels. *Transportation Res F Traffic Psychol Behav* (2007) 11(1):61–74. doi:10.1016/j.trf.2007.07.001

6. Pachamanova A, Dessislava P. Optimization of the light distribution of luminaries for tunnel and street lighting. *Eng Optimization* (2008) 40(1):47–65. doi:10.1080/03052150701591160

7. Mashimo H. State of the road tunnel safety technology in Japan. *Tunneling Underground Space Tech* (2002) 17(2):145–52. doi:10.1016/s0886-7798(02)00017-2

8. Auboyer A, Andersen V, Wybo J-L. State-of-the-art road tunnel safety. *Int J Emerg Manage* (2007) 4(4):610–29. doi:10.1504/ijem.2007.015733

9. Wang S. The experience and enlightenment of the development of smart highways in the UK. *Auto Saf* (2020) 276(12):104–109. doi:10.3969/j.issn.1006-6713.2020.12.029

10. Wu J. Yanqing to Chongli Highway radar road condition sensing system. *China ITS J* (2021) 1(9):105–7. doi:10.13439/j.cnki.itsc.2021.01.009

11. Zhang J, Li B, Wang X. Design of architecture and development path of intelligent highway. *J Highw Transportation Res Develop* (2018) 35(1):88–94. doi:10.3969/j.issn.1002-0268.2018.01.012

12. Xia Y, Yan C, Wang X, Song X. Intelligent transportation information physical fusion cloud control system. *Acta Automatica Sinica* (2019) 45(1):132–42. doi:10.16383/j.aas.c180370

13. Cen Y, Song X, Wang D, Sun L, Liu N. Construction of smart highways technology system. *J Highw Transportation Res Develop* (2020) 37(7):111–21. doi:10.3969/j.issn.1002-0268.2020.07.015

14. Wang S, Zu H, Fu J, Ruan Z, Li M. Exploring the smart highway. *China ITS J* (2017) S1:10–7. doi:10.13439/j.cnki.itsc.2017.S1.001

15. Fu L, Xie S, Fan Z. The application of fuzzy matter-element model in the safety evaluation of highway tunnels operation. *Technology Highw Transport* (2015) 117(2):122–6. doi:10.13607/j.cnki.gljt.2015.02.027

16. Yuan Y, Cao B. Analysis on information infrastructure in the intelligent high-speed + Internet environment. *China ITS J* (2019) 11(237):139–41. doi:10.13439/j.cnki.itsc.2019.11.019

17. Du Y, Liu C, Wu D, Zhao C. Next-generation intelligent highway system architecture design. *China J Highw Transport* (2022) 35(04):203–14. doi:10.19721/j.cnki.1001-7372.2022.04.017

18. Li Y, Liu C, Yue G, Gao Q, Du Y. Deep learning-based pavement subsurface distress detection via ground penetrating radar data. *Automation in Construction* (2022) 142:104516. doi:10.1016/j.autcon.2022.104516

19. Li Y, Liu C, Gao Q, Wu D, Li F, Du Y. ConTrack distress dataset: a continuous observation for pavement deterioration spatio-temporal analysis. *IEEE Trans Intell Transportation Syst* (2022) 23(12):25004–17. doi:10.1109/tits.2022.3201968

20. Huval B, Wang T, Tandon S, Kiske J, Song W, Pazhayampallil J, et al. An empirical evaluation of deep learning on highway driving. *Computer Sci V3* (2015) 1–3. doi:10.48550/arXiv.1504.01716

21. Iwasaki Y, Misumi M, Nakamiya T Robust vehicle detection under various environmental conditions using an infrared thermal camera and its application to road traffic flow monitoring. *Sensors* (2013) 13(6) 7756–73. doi:10.3390/s130607756

22. Wang M, Wang R. Application prospects of digital twin technology in the field of intelligent highways. *China ITS J* (2019) S1:34–5. doi:10.13439/j.cnki.itsc.2022.S1.008

23. Gongsong D, Wang L, Wang S. Energy consumption monitoring index system for highway tunnel operation period. *Chin J Underground Space Eng* (2020) 16(S1):407–12.

24. Wu X, Deng T, Chen B, Zeng T, Chen H, Zhang K. Research on compressed sensing of big data for structural health monitoring system of operating tunnel. *Tunnel Construction* (2021) 41(4):674–83. doi:10.3973/j.issn.2096-4498.2021.04.019

25. Xiao H. Research on the long-term operation structure health monitoring system of Shanghai dalian Road tunnel. *China Municipal Eng* (2016) 183(1):75–77. doi:10.3969/j.issn.1004-4655.2016.01.024

26. Yu S, Chen Y, Song L, Xuan Z, Li Y. Modelling and mitigating secondary crash risk for serial tunnels on freeway via lighting-related microscopic traffic model with inter-lane dependency. *Int J En-vironmental Res Public Health* (2023) 20:3066. doi:10.3390/ijerph20043066

27. Yu S, Shi L, Zhang L, Liu Z, Tu Y. A solar optical reflection lighting system for threshold zone of short tunnels: theory and practice. *Tunnelling Underground Space Tech* (2023) 131:104839. doi:10.1016/j.tust.2022.104839

28. Yu S, Zhao C, Song L, Li Y, Du Y. Understanding traffic bottlenecks of long freeway tunnels based on a novel location-dependent lighting-related car-following model. *Tunnelling and Underground Space Technol-ogy* (2023) 136:105098. doi:10.1016/j.tust.2023.105098

29. Sun J, Qi G, Mazur N, Zhu Z. Structural scheduling of transient control under energy storage systems by sparse-promoting reinforcement learning. *IEEE Trans Ind Inform* (2022) 18(2):744–56. doi:10.1109/TII.2021.3084139

30. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf Fusion* (2023) 91:376–87. doi:10.1016/j.inffus.2022.10.022

31. Zhu Z, Lei Y, Qi G, Chai Y, Mazur N, An Y, et al. A review of the application of deep learning in intelligent fault diagnosis of rotating machinery. *Measurement* (2023) 206:112346. doi:10.1016/j.measurement.2022.112346

32. Liu Y, Zhang Y, Liu G, Zhang H, Jin L, Fan Z, et al. Design of a traffic flow data collection system based on LiDAR-Vision Integrated Machine. *Comp Meas Control* (2021) 1–9. doi:10.16526/j.cnki.11-4762/tp.2022.03.027

33. Lu X. Real-time and spatial traffic perception system based on LiDAR. *China ITS J* (2021) 253(2):122–5. doi:10.13439/j.cnki.itsc.2021.02.015

34. Wang S, Zhang J, Zhao C, Wang W. Discussion on the application of big data technology in highway tunnel engineering. *Highway* (2017) 62(8):166–73.

35. Hong K. Development and thinking of tunnels and underground engineering in China in Recent 2 Years (From 2017 to 2018). *Tunnel Construction* (2019) 39(5):710–23.

36. Yang C, Nie Q. Time series decomposition and machine learning fusion model for PM2.5 prediction. *J Saf Environ* (2023) 1–11. doi:10.13637/j.issn.1009-6094.2022.1616

37. Li M, Ning D, Guo J. Attention-based CNN-LSTM model and its application. *Comp Eng Appl* (2019) 55(13):29–27.

38. Wang S, Qiao Z, Fu J, Li M. The connotation and architecture of smart highways. *Highway* (2017) 62(12):170–175.

39. Zeng L, Wang S, He X, Ruan Z. The concept, architecture, and key technologies of intelligent highway tunnels. *Mod Tunnelling Tech* (2016) 53(04):1–8. doi:10.13807/j.cnki.mtt.2016.04.001

40. Li P, Mao Y, Wang S, Fu L, Yang X. Research on multi-scale early warning system for safe operation of highway tunnels. *China ITS J* (2021) 256(S1):60–4. doi:10.13439/j.cnki.itsc.2021.S1.014

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 13 February 2024
DOI 10.3389/fphy.2024.1287050

OPEN ACCESS

EDITED BY
Huafeng Li,
Kunming University of Science and Technology,
China

REVIEWED BY
Feng Xu,
Nanjing University of Posts and
Telecommunications, China
Jinxing Li,
Harbin Institute of Technology, Shenzhen,
China

*CORRESPONDENCE
Liye Mei,
✉ liyemei@whu.edu.cn

RECEIVED 01 September 2023
ACCEPTED 30 January 2024
PUBLISHED 13 February 2024

CITATION
Chen S and Mei L (2024), Structure similarity virtual map generation network for optical and SAR image matching.
Front. Phys. 12:1287050.
doi: 10.3389/fphy.2024.1287050

Structure similarity virtual map generation network for optical and SAR image matching

Shiwei Chen[1] and Liye Mei[2,3]*

[1]Department of Automation, Rocket Force University of Engineering, Xi'an, China, [2]School of Computer Science, Hubei University of Technology, Wuhan, China, [3]The Institute of Technological Sciences, Wuhan University, Wuhan, China

Introduction: Optical and SAR image matching is one of the fields within multi-sensor imaging and fusion. It is crucial for various applications such as disaster response, environmental monitoring, and urban planning, as it enables comprehensive and accurate analysis by combining the visual information of optical images with the penetrating capability of SAR images. However, the differences in imaging mechanisms between optical and SAR images result in significant nonlinear radiation distortion. Especially for SAR images, which are affected by speckle noises, resulting in low resolution and blurry edge structures, making optical and SAR image matching difficult and challenging. The key to successful matching lies in reducing modal differences and extracting similarity information from the images.

Method: In light of this, we propose a structure similarity virtual map generation network (SVGNet) to address the task of optical and SAR image matching. The core innovation of this paper is that we take inspiration from the concept of image generation, to handle the predicament of image matching between different modalities. Firstly, we introduce the Attention U-Net as a generator to decouple and characterize optical images. And then, SAR images are consistently converted into optical images with similar textures and structures. At the same time, using the structural similarity (SSIM) to constrain structural spatial information to improve the quality of generated images. Secondly, a conditional generative adversarial network is employed to further guide the image generation process. By combining synthesized SAR images and their corresponding optical images in a dual channel, we can enhance prior information. This combined data is then fed into the discriminator to determine whether the images are true or false, guiding the generator to optimize feature learning. Finally, we employ least squares loss (LSGAN) to stabilize the training of the generative adversarial network.

Results and Discussion: Experiments have demonstrated that the SVGNet proposed in this paper is capable of effectively reducing modal differences, and it increases the matching success rate. Compared to direct image matching, using image generation ideas results in a matching accuracy improvement of more than twice.

KEYWORDS

structural similarity, multi-sensor, virtual map, image matching, deep learning, generative adversarial networks, SAR images

1 Introduction

With the advancement of satellite remote sensing technology [1], the means of data acquisition are constantly being enriched. How to effectively integrate multi-sensor, high-resolution, multi-spectral, and multi-temporal remote sensing data for fusion processing has become a hot and key research topic in the field of remote sensing at present. Multi-source image matching [2, 3], especially the matching between optical and SAR images [4, 5], is one of the core problems that urgently needs to be solved. However, due to the completely different imaging mechanisms, there are radiation anomalies, geometric differences, and scale differences between optical and SAR images. This increases the difficulty of image matching and makes SAR and optical image matching an international challenge.

Currently, multi-modal image matching can be categorized into three main types: region-based matching, feature-based matching, as well as deep learning-based matching. Region-based image matching places emphasis on comparing local regions in the images by calculating grayscale information and establishing correlation signals. Common similarity measurement functions [6] include SSD, NCC, MI, and PC. However, region-based matching methods are sensitive to nonlinear grayscale distortions, making them less suitable for multi-modal image matching. Feature-based matching methods [7] extract common features from reference and target images and establish correspondences to determine the transformation model parameters for matching. These features include region features, line features (extracted from edges and texture information) and point features. Point features are the most extensively studied, involving the extraction of key points with certain invariance properties and their description using specific descriptors. Common methods for point feature extraction include Harris corner detection, SIFT [8], and SURF [9]. Researchers have also proposed geometric structure-based feature [10] descriptors like HOPC, CFOG and RIFT [11] to meet the requirements of multi-modal images. Feature-based matching methods provide higher-level information beyond grayscale and offer adaptability to grayscale variations, image deformations, and occlusions, thereby broadening the application scope of image matching techniques.

The popular deep learning methods in recent years are mainly divided into single-loop deep neural network and end-to-end deep networks. Single-loop deep neural networks include D2-Net, Superglue, and so on. End-to-end deep networks include MUNIT-based multi-modal image matching, Dual-Attention Networks for multi-modal image matching, Cross-Modal Feature Fusion and generative adversarial networks (GAN). Furthermore, the basic ideas of style transfer methods [12] and end-to-end patterns are the same. By utilizing deep learning networks [13] to obtain optical image features, replicate attributes originating from SAR data onto optical representations, and then match them using traditional methods, such as SIFT, SURF, and RIFT. The goal of these approaches is to maintain consistency [14] between the transformed SAR images and the original images, followed by feature matching with traditional methods. These methods require further research on the depth matching framework, the loss function [15], and training strategies with the intention of improving matching performance for heterogeneous remote sensing image matching.

Consequently, the pursuit of efficacious strategies to mitigate feature matching discrepancies bears substantial practical research implications. This is done by enhancing consistency between generated and original images, and achieving robust matching of heterogeneous images. In light of this, we study style transfer methods and perform feature transformation on SAR images. This is to ensure that the traits of the generated SAR image align with those of the corresponding optical image, thereby optimizing the matching of heterogeneous images.

In this paper, we propose the SVGNet to seek effective methods for reducing modal differences. This framework leverages Conditional Generative Adversarial Network (CGAN), Attention U-Net, SSIM, and LSGAN to generate virtual maps and optimize multi-modal image matching. Specifically, for feature learning without the need for additional supervision, we employ Attention U-Net with attention gates that automatically focus on salient feature regions during feature learning. Therefore, we utilize Attention U-Net as the generator to extract image features. Additionally, we transform the task of multi-modal image matching into the task of reducing modality differences, for which CGAN is employed to generate virtual maps and minimize modality disparities. By incorporating conditional constraints, CGAN controls the details of image generation to achieve desired effects, making this model exceptionally effective. Finally, to optimize the overall training performance of the generative model and improve the realism of generated images, we utilize SSIM to constrain spatial information and enhance image quality. Simultaneously, LSGAN is employed to stabilize SVGNet training. To validate the effectiveness of our proposed method, we conduct extensive experiments to demonstrate SVGNet's superiority over other generative adversarial networks. We also demonstrate the quality of our generated virtual maps. The results indicate that SVGNet has advantages in the direction of multi-modal image matching. The major contributions of this paper can be summarized as follows:

1. We introduce SVGNet, an innovative approach to meet the challenges of optical and SAR image matching.
2. We employ CGAN to reduce dissimilarities between matched images and generate superior-quality images specifically tailored for matching task.
3. We adopt an Attention U-Net in a decoder module, to extract and learn features from optical images to better focus on relevant regions of the images.
4. We utilize SSIM and LSGAN losses to amplify the model's optimization performance and foster training stability.
5. We conduct extensive experiments to study in detail the high-quality impact of the generating virtual maps and the superior performance of the network.

The results show that the SVGNet proposed in this paper shows superiority in the quantitative analysis of optical and SAR image matching.

2 Related work

In the most recent years, deep learning [16] has gained attention and accomplished significant advancements in fields like visual cognition and natural language understanding. Researchers have proposed deep learning-based methods [17] for multi-source image matching. These methods can be categorized into two aspects:

2.1 Single-loop deep neural network

Single-loop deep neural network, which only replaces some matching links, is often more flexible and can meet different needs by combining other advantageous structures to build a complete matching model. Numerous scholars harness the power of deep learning to meticulously detect a significantly enhanced and dependable set of salient critical points from images, adeptly acquiring the principal orientation or predominant scale for each individual feature point, along with refining more discriminative and correspondingly matchable feature descriptors. At the beginning, Dusmanu et al. [18] innovatively constructed the network structure D2-Net, which integrates detection features and feature description. The key points are extracted by slicing the feature map, using convolutional neural networks (CNN) to calculate the descriptors. By improving D2-net, MA et al. [19] demonstrated CMM-Net and applied it to multi-modal image matching. This method used dynamic adaptive Euclidian distance threshold and RANSAC algorithm to eliminate the wrong matching points and showed excellent matching effect in the image matching of alien remote sensing images. Hao et al. [20] designed a multi-level semantic extractor to extract rich and diverse semantic features from real images to effectively guide sample generation. Ma et al. [21] explored a matching method integrating deep learning with conventional local features from rough to fine, extracted deep features through CNN for rough matching, and then adjusted the rough matching results by combining more accurate local features, so as to produce more stable matching results. To learn descriptor representations of multimodal image blocks, Zhang et al. [22] used maximum positive sample and negative sample feature distances as loss functions in their full-convolutional neural network (FCN) built upon the Siamese network structure. Subsequently, Li et al. [10] presented a rotation-invariant multi-modal image matching method grounded in deep learning jointly with Gaussian features. A neural architecture referred to as RotNET underwent training to forecast the rotational interrelationship among images. Subsequently, the alignment of two images was achieved through the establishment of gradient-oriented Gaussian pyramid features (GPOG). Some scholars also use deep learning to learn more reliable similarity measurement criteria and gross error elimination among descriptors. Sarlin et al. [23] designed a representative network superglue for feature matching and gross error elimination. This neural framework approaches the challenge of feature matching by framing it as the task of addressing the differentiable optimal transport quandary. Recurrent neural network (RNN) is constructed to solve this problem. Ma et al. [24] employed deep learning techniques to devise a gross error elimination network, denoted as LMR, bearing resemblance to the RANSAC algorithm. This approach translated the task of gross error elimination into a binary classification paradigm. The deep learning network was harnessed to assess the validity of each initial match pair, culminating in the successful mitigation of gross errors. These approaches leverage the robust deep feature extraction proficiency and the adeptness in high-dimensional feature representation offered by deep learning methodologies. By training a single network to replace a certain link in multi-modal image matching, these methods are combined with others to construct a comprehensive multi-modal image matching model, which has greater flexibility in use.

2.2 End-to-end deep neural network

Devise an end-to-end matching network directly predicated upon the principles of deep learning. The framework consists of three neural network structures for feature extraction, feature matching, and outlier removal, which provide excellent matching results pertaining to images obtained by optical and SAR techniques. In Hughes et al. [25], a neural network based algorithm for automatically matching multi-scale and multi-modal images has been developed, consisting of three neural network structures, corresponding to feature space extraction, matching based on feature space correlation functions, and outlier elimination, respectively. The matching effect for optical and SAR images is excellent. The KCG-GAN algorithm, as outlined in [26], incorporates K-means segmentation as an input modality for the image synthesis process. Through the imposition of spatial information synthesis constraints, it enhances the fidelity of synthesized imagery, and its application encompasses the realm of SAR and optical image alignment. Nevertheless, owing to the higher requirements of multi-modal image training data sets, and the complexity of imaging differences, mixed noise, and regional gray level differences between images. Sun et al. [27] described the LoFTR matching method of Canonical, which detects, describes, and matches image features on a coarse-grained basis, before refinement of the intensive subpixel matching on a fine-grained basis. Moreover, the Transformer model employs self-attention and cross-attention mechanisms as foundational components for generating feature descriptors from a pair of images. End-to-end networks can also be used to preprocess images, using techniques such as image synthesis and style transfer. Based on the imaging characteristics of different modal images, transform the style of images in different modalities, and used to expand the multi-modal image dataset or directly convert it into the same modal image form for matching.

3 Methods

3.1 Network architecture

Our objective is to achieve a better matching effect between SAR images and optical images, and the key lies in reducing modal differences between them. As shown in Figure 1, the red box represents our proposed SVGNet based on GAN. By introducing the concept of style transfer, the network generates novel images that bridge the gap between single-mode and multi-modal datasets, showcasing the process of image-to-image conversion. The fundamental idea of SVGNet is to train the generative model through adversarial training. In other words, through mutual competition and learning, the generation model and the discrimination model are constantly improved to achieve the optimal state.

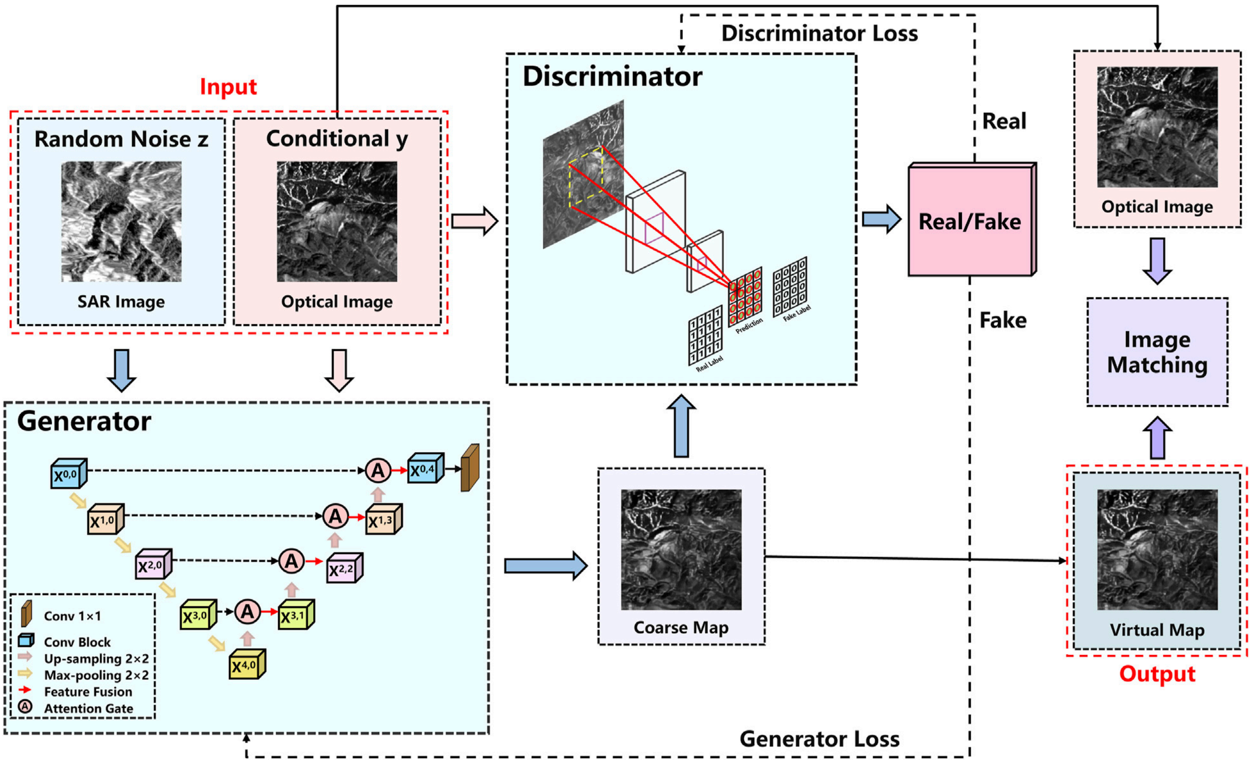


FIGURE 1
The architecture of optical and SAR image matching method based on SVGNet.

However, the unrestricted nature of GANs, lacking prior modeling, poses challenges in controlling them effectively for large-scale images with numerous pixels. To tackle this challenge, our proposition involves the incorporation of CGAN into the framework. According to Figure 1, the condition variable we use in this paper is the original optical image. By connecting the real optical image and its label, we can determine whether an image is a "real" image or a "fake" image. A fake label is generated as a condition for generating the optical image using the true optical image.

The proposed SVGNet has the following four improvements: (1) To enhance the network's training capability and achieve desired data generation, we introduce CGAN and modify the unsupervised GAN [28] to a supervised GAN. This modification involves incorporating conditional information and adjusting the generator and discriminator. (2) This network uses Attention U-Net [29], which provides a more flexible structure, higher-quality image generation, and better preservation of semantic information than KCG-GAN. Optical images serve as conditional information, while the original SAR image labels serve as random noise. These two factors are fed into the generator in order to generate initial coarse maps, which then guide the optimization of feature learning. (3) On the other hand, the discriminator utilizes a fully convolutional neural network to ensure training stability and evaluate the authenticity of generated images. Optical images serve as conditional information, and the coarse map labels generated by the discriminator are used to evaluate authenticity. The discriminator plays a crucial role in determining the authenticity of refined maps. (4) Additionally, the losses of the generator and discriminator are computed. The SSIM is applied throughout the training process to enhance spatial constraints and improve image quality. Moreover, the training utilizes LSGAN to stabilize SVGNet. Once the losses reach saturation and a certain number of iterations are reached, a virtual map is generated.

With the generated virtual maps, we can perform better image matching. Below, we will discuss in more detail the specific modules and loss functions used in SVGNet.

3.1.1 Generative network

We propose to generate virtual maps to promote more efficient matching of optical and SAR images. Thus, in the generation network, it is essential for the generator to accurately and effectively extract the features of optical images. Furthermore, high-resolution input grids to high-resolution output grids are the hallmark of image-to-image transformation challenges. Additionally, the input and output appear differently on the surface, but they are both rendered with the same underlying structure. Consequently, the input and output structures are roughly aligned. We formulate the generator architecture with these considerations at its core. Therefore, we use Attention U-Net as a generator, as shown in the Generator module in Figure 1, which has image reconstruction capability and an attention mechanism. First, the proposed network consists of an encoder and a decoder. Specifically, the encoder learns the potential features of the original optical images, while the decoder is responsible for reconstructing from the low-level feature to the high-level feature to obtain the generated optical images.

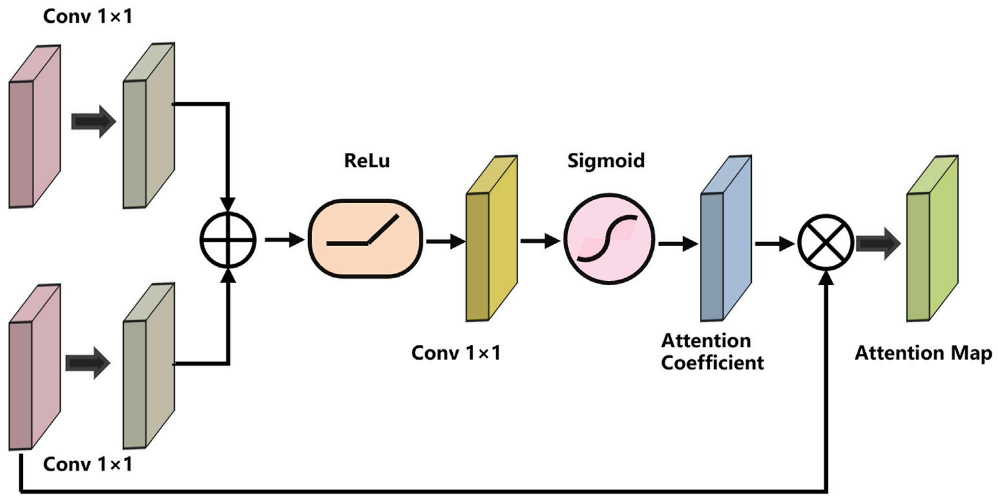


FIGURE 2
The structure of the Attention Gate (AG).

To simplify the description of the network, we refer the convolution layer [30], Batch Norm layer [31], and Rectified Linear Unit [32] as Conv, BN, and ReLu respectively. The structure of Attention U-Net can be seen in the generator module in Figure 1. The output of the node $X^{i,j}$, which is denoted as $x^{i,j}$, is defined as Eq. 1:

$$x^{i,j} = \begin{cases} C\left(D\left(x^{i-1,j}\right)\right), \ j = 0 \\ C\left(\left[A\left(x^{i,j-1}, U\left(x^{i+1,j-1}\right), U\left(x^{i+1,j-1}\right)\right)\right]\right), j > 0 \end{cases} \tag{1}$$

In the equation, the functions $C(\cdot)$, $D(\cdot)$, $U(\cdot)$, $A(\cdot)$ and $[\cdot]$ denote the convolution, down sampling, up sampling, AG, and concatenation operations, respectively. The convolutional block consists of two Conv-BN-ReLU layers, each employing a filter size of 3 × 3, a padding of 1, and a stride of 1. This configuration is strategically designed to ensure the output feature map preserves the identical dimensions as the input. The downward arrows indicate a 2 × 2 max-pooling layer, and the upward arrows indicate 2 × 2 up-sampling, aiming at decoding low level feature map to acquire a high-resolution feature map. Second, to address the challenge of image consistency, an attention module (Attention Gate, AG) is introduced to the U-Net architecture as depicted in Figure 2. It is aimed at highlighting significant features by skipping connections, extracting information from rough scale to distinguish irrelevant features from noise, and letting the value of irrelevant regions be suppressed and the value of target regions become larger. By generating a gated signal, AG effectively modulates the significance of features across diverse spatial locales. This signal serves to prioritize attention on salient features deemed valuable for tasks related to phase recovery, while concurrently dampening the influence of extraneous regions within the input image. Intuitively, it inserts an AG in each skip connection, which concatenates the same-level $x^{i,j-1}$ feature map with the up-sampled feature map $U(x^{i+1,j-1})$ as input. Then, through ReLU and Sigmoid operations, the attention coefficient map is obtained. Finally, the inner product of the attention coefficient map and the up-sampled feature map is used to obtain the attention map. Consequently, the network will allocate heightened focus toward the attributes inherent in the optical image.

In general, the Attention U-Net network is used in this paper because it is capable of extracting image details well and retaining image information on different scales. The AG of Attention U-net improves the discernment and precision of the dense feature prediction model and improves the prediction accuracy. CGAN can effectively transform both deep feature information in the image and deep feature information that cannot be transformed. Attention U-Net encodes 256 × 256 input SAR images in the coded down-sampling and then decodes and up-sampling after the down-sampling is completed. The output image is still 256 × 256 in size.

3.1.2 Discriminant network

Compared to the original GAN discriminator, the Markov discriminator (Markovan Discriminator) is one of the discriminators in CycleGAN. As shown in Figure 1, the discriminant network is not implemented by utilizing various convolution layers that are then input into the connection layer or activation function, but by using a sliding window approach to determine whether individual patches are genuine and authentic. By upholding local coherence, this approach enables the generative network to discern finer-grained information from its contextual surroundings.

This paper divides the discriminant images into $N \times N$ patches as input to the discriminant network. Every element in the output matrix indicates the likelihood of the corresponding image patch being authentic or synthetically generated. By analyzing the structural features of each patch in the image, the network can better process the high-frequency information part of the image.

3.2 Loss function

The loss functions used in this paper include the SSIM and LAGAN loss functions, which will be introduced in detail below.

(1) SSIM

Based on the network framework of CGAN, the algorithm replaces random noise as input. For supervised segmentation, we adopt SSIM

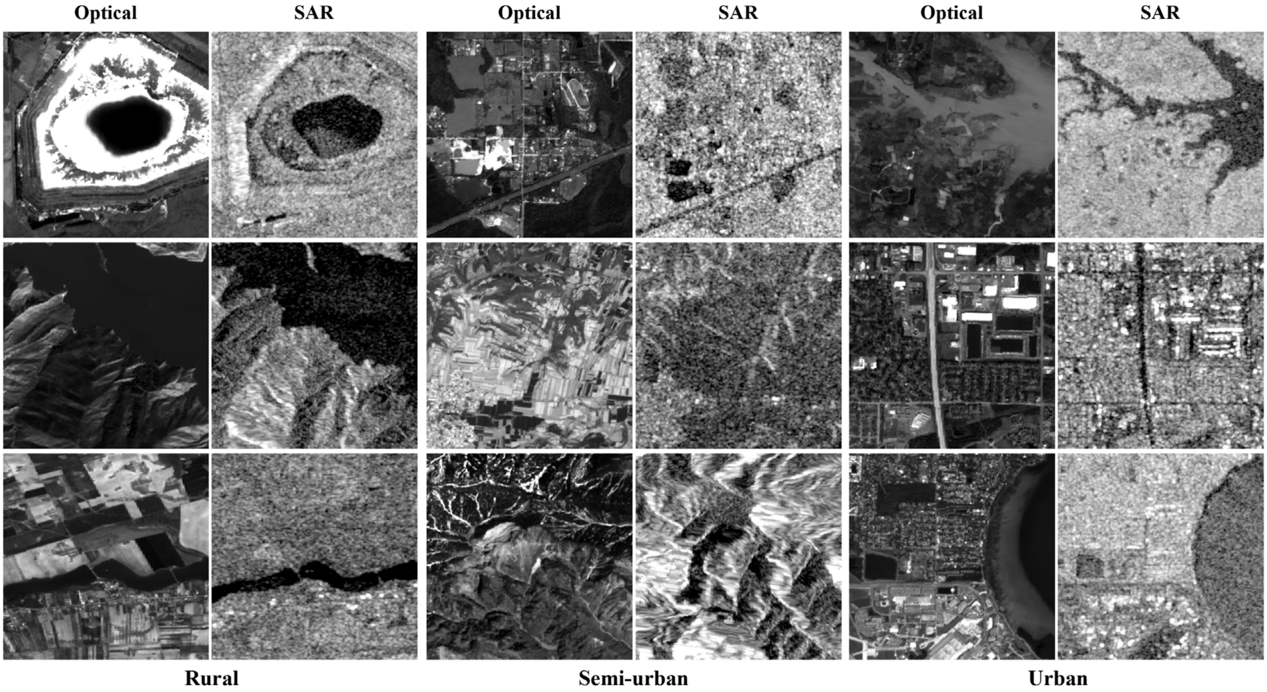


FIGURE 3
SEN1-2 data set. Three different scenarios: Rural, Semi-urban and Urban; Each group has optical images on the left and SAR images on the right.

loss [33] with the objective of making the segmentation map as close to the ground truth as possible. SSIM can be defined by Eq. 2:

$$L_{SSIM}(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{2}$$

Where x, y denote the phase images restoration results and the ground truth, μ_x, μ_y and $\sigma_x^2 + \sigma_y^2$ are the mean and the deviations of the image respectively, σ_{xy} is the covariance for the x, y and C_1, C_2 are small constants.

(2) LSGAN

Regular GAN loss can suffer from model collapse and is notoriously difficult to converge.

Due to the fact that LSGANs are more stable and have been shown in previous experiments to be capable of achieving better segmentation results, we adopt them as the loss function in our work [34], since they are more stable and have been shown to achieve better segmentation results. It is defined by Eq. 3:

$$\begin{aligned} L_{LSGAN}(D) = {} & E_{i,y\sim P_{data(i,y)}}\left[\left(D(i,y)-1\right)^2\right] \\ & + E_{i\sim P_{data(i)}}\left[\left(D(i,G(i))\right)^2\right] \end{aligned} \tag{3}$$

Furthermore, the adversarial learning process can be notably enhanced by employing LSGAN, as expounded in Eq. 4 below:

$$L_{LSGAN}(G) = E_{i\sim P_{data(i)}}\left[\left(D(i,G(i))-1\right)^2\right] \tag{4}$$

In the Eqs 3, 4, i is the input and y is the ground truth.

(3) Final loss function

The objective function for SVGNet is defined by Eq. 5:

$$\begin{gathered} \min_D L(G) = L_{LSGAN}(D) \\ \min_G L(G) = L_{LSGAN}(G) + \lambda L_{SSIM} \end{gathered} \tag{5}$$

where λ governs the relative importance of the two objective functions. As a matter of experience, we set λ to 10 in our work.

4 Experiment and analysis

4.1 Datasets

This paper utilizes the widely-used SEN1-2 dataset [35], which provides a comprehensive collection of aligned Sentinel 1 SAR and Sentinel 2 optical images. In this context, the dataset consists of 282,384 image pairs with a resolution of 256 pixels and an 8-bit depth. It encompasses diverse geographical regions and countries, capturing various features such as cities, agricultural land, forests, mountains, and water bodies. The following three scenarios were selected for a comprehensive evaluation: rural (300 image pairs), semi-urban (300 image pairs), and urban (300 image pairs). The trained model then applies style transfer to the test set, generating images depicting cities, towns, and countryside landscapes. The dataset allows for a clear separation between training and testing data, enabling an unbiased performance evaluation. Notably, this dataset has been extensively used in deep learning-based alignment studies for SAR and optical images. Figure 3 provides representative samples from the dataset. There are three different scenarios: Rural, Semi-urban and Urban. Each group has optical images on the left and SAR images on the right.

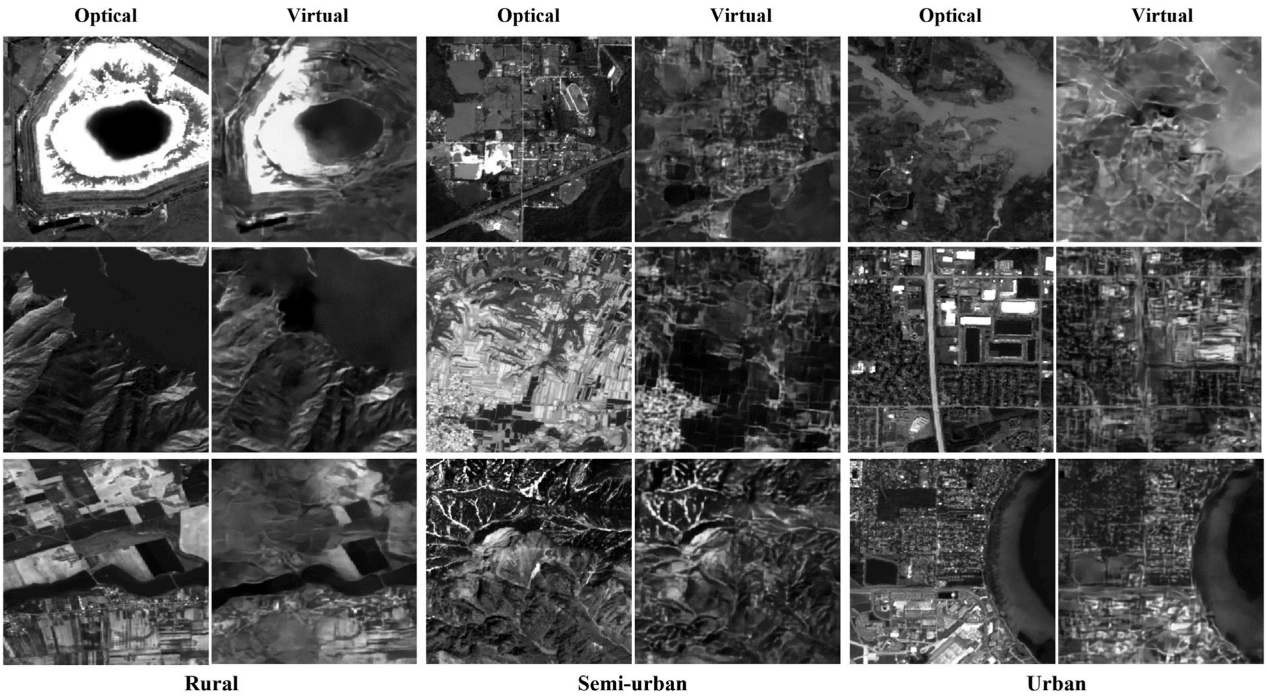


FIGURE 4
The performance of the virtual maps. There are three types of scenes: rural, semi-urban and urban, each of which shows 3 groups of images. In each group, the left column features the generated virtual maps, while the right column displays the corresponding original optical images.

4.2 Experimental details

4.2.1 Evaluation metrics

In this paper, we will analyze the quality of virtual maps and the effectiveness of image matching. Therefore, three metrics are selected for evaluating the effectiveness of image matching: NCM (Number of Correct Matching points), Matching success rate and RMSE (Root Mean Square Error) [16]. Evaluation Metrics for effectiveness of image matching are defined as follows:

(1) Number of Correct Matching points (NCM) indicates the number of feature points correctly matched between two images. Consequently, the higher the NCM value, the more accurate the matching results are.

(2) Matching Success Rate (MSR) is known as matching accuracy. It is a performance metric used to evaluate the accuracy of image matching algorithms. A higher matching correctness rate indicates a more reliable and accurate matching result, which is due to the algorithm's ability to correctly identify and match corresponding points across the images. It is computed through the division of NCM by the total number of matched points. The formula can be defined as follows in Eq. 6:

$$MSR = \frac{NCM}{Total\ number\ of\ matching\ points} * 100\% \tag{6}$$

(3) Root Mean Square Error (RMSE) means that the point coordinates of the same label in the benchmark image and the prepared matching image are labelled as (x_i, y_i) and $(x_i^{'}, y_i^{'})$ respectively. S represents the number of the points with the same label selected; $(x_i^{'}, y_i^{'})$ is the coordinate of the $i\,th$ prepared matching image pair of the same label (x_i, y_i) after the matching correspondence conversion. RMSE is defined as follows in Eq. 7:

$$RMSE = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(x_i - x_i^{'}\right)^2 + \left(y_i - y_i^{'}\right)^2} \tag{7}$$

4.2.2 Parameter settings

All experimental endeavors are executed within the PyTorch framework, renowned for its adeptness in high-performance computation. For computation, a sole NVIDIA Tesla A100 GPU is deployed, replete with a GPU memory capacity of 80 GB. For the duration of the model's training period, a batch size of 8 is employed, with each model undergoing a maximum of 1000 training epochs. The optimization process is facilitated by Adam, chosen for its efficacy, and initialized with a learning rate of 0.002 to circumvent issues tied to insufficient learning weight. To preclude overfitting during the training process, the early stopping technique is judiciously incorporated.

4.3 Image generation results and analysis

Visually, it is observable that the radiation difference between the SAR generated image and the original optical image is reduced. For certain images, such as the bottom row image in the Semi-urban group of Figure 4, the virtual maps generated by our SVGNet are almost identical to the optical

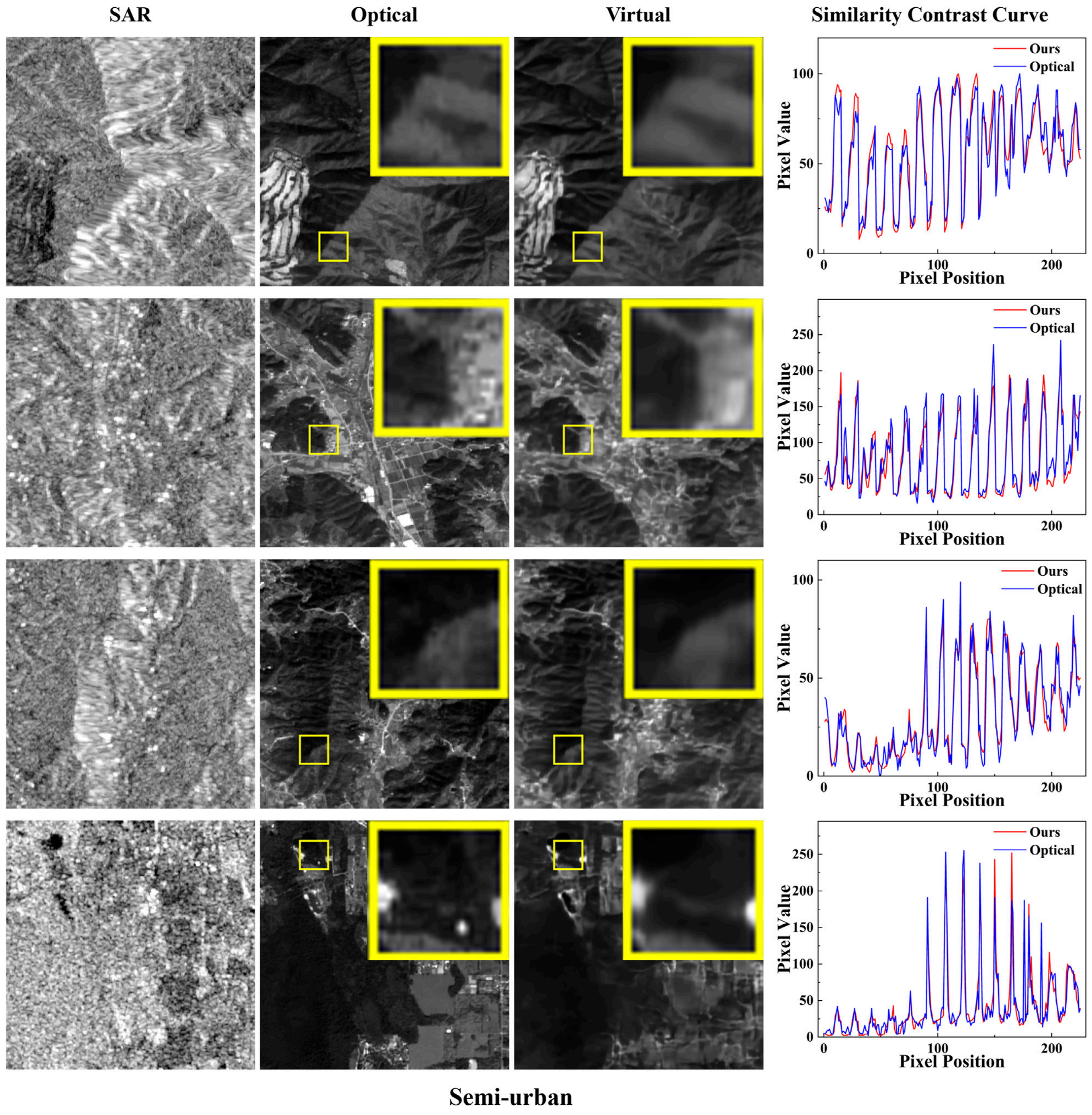


FIGURE 5
Comparison of virtual maps and optical images in the semi-urban group. From left to right are the corresponding SAR image, optical image, generated virtual maps and pixel contrast curve of the virtual maps; Mark the randomly selected test area with a pixel size of 30 × 30 with a yellow box and place it at the image's upper right corner; Contrast curve of pixel values (red: the virtual maps generated by us; blue: corresponding optical image; horizontal axis: pixel position; vertical axis: corresponding pixel value).

images, with clear edge textures and nearly identical shapes. The grayscale is similar, the size, shape and relative position of the objects are almost the same. In virtual maps, the texture and fine features of the original optical image can be preserved. Several areas have been cut for enlargement display and quantitative analysis has been performed in order to better display the generation effect.

For the purpose of quantitative analysis, we randomly selected 4 groups of data separately from the semi-urban and urban scenarios for testing. After that, random pixel values are extracted from rows and columns and drawn into one dimension for each group of graphs. To compare the pixel values of corresponding positions, the curve of pixel values of the two graphs is drawn on a graph, as shown in Figures 5, 6. In the result graph, it can be seen that the curve fitting degree of the pixel values is extremely high, which indicates that the virtual maps generated by SVGNet method are very similar to the optical original image, and the effect is truly remarkable.

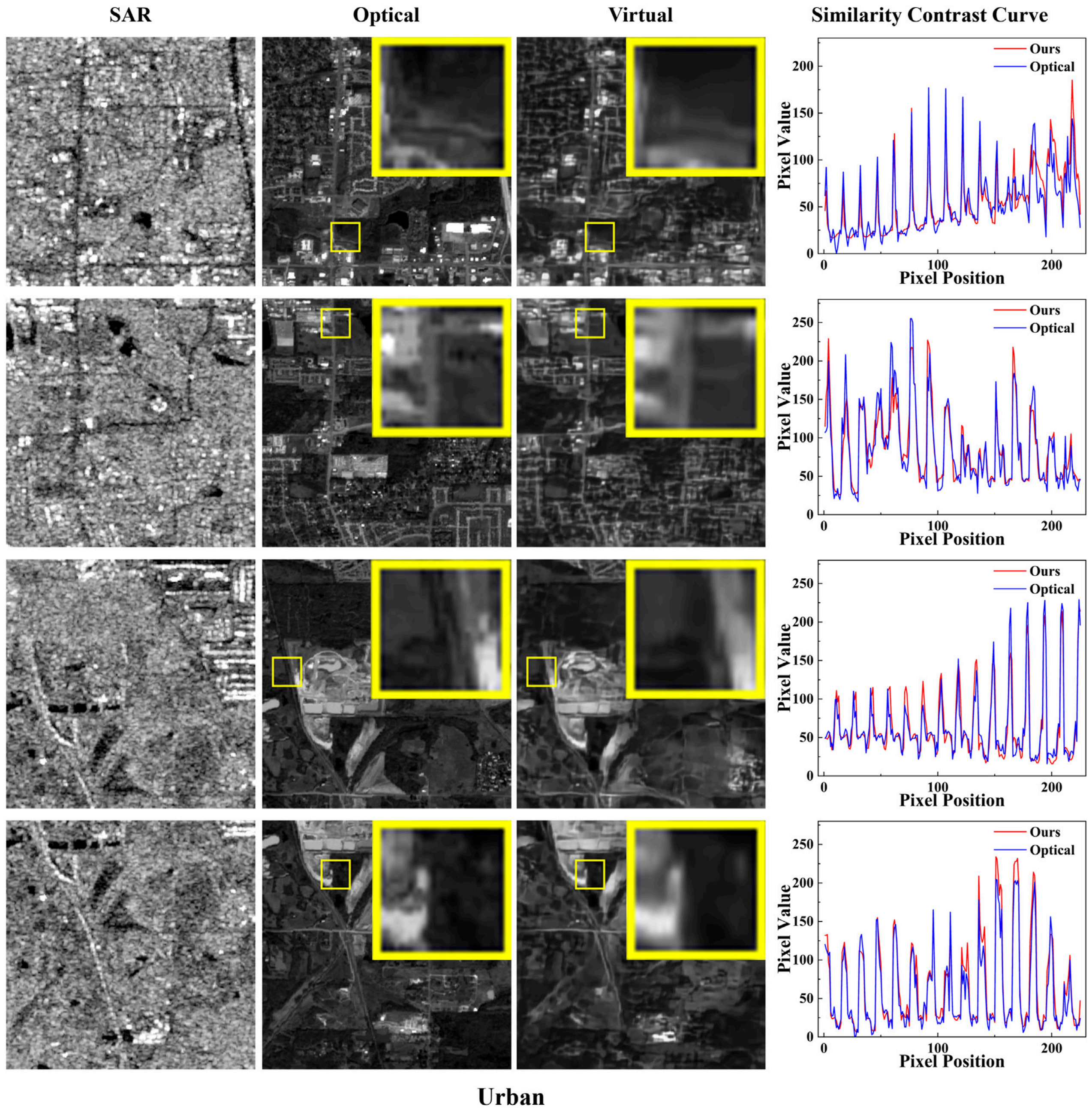


FIGURE 6
Comparison of virtual maps and optical images in the Urban group. From left to right are the corresponding SAR image, optical image, generated virtual maps and pixel contrast curve of the virtual maps; Mark the randomly selected test area with a pixel size of 30 × 30 with a yellow box and place it at the image's upper right corner; Contrast curve of pixel values (red: the virtual maps generated by us; blue: corresponding optical image; horizontal axis: pixel position; vertical axis: corresponding pixel value).

4.4 Matching effect comparison and analysis

We compare SVGNet for image matching from two perspectives in this paper in order to evaluate its effectiveness: (1) Comparing the generated adversarial network between KCG-GAN and SVGNet in this paper, the matching method adopts the traditional RIFT algorithm; (2) Comparison of matching methods. This paper compares the proposed method to three baseline methods, including LoFTR, D2-Net, and Superglue. LoFTR is an end-to-end deep network, while D2-Net and Superglue are single-loop networks. Initially, LoFTR establishes coarse-grained image feature detection and matching, and then refines subpixel-level intensive matching to refine the results. Furthermore, Transformer uses both self-attention layers in order to obtain feature descriptors for two images, and it also utilizes mutual attention layers in order to do so. D2-Net innovatively constructs a network structure integrating detection features and feature descriptions. Descriptors were calculated by slicing CNN feature maps, and then key points are extracted by calculating descriptors. Superglue solves this problem by treating the feature matching problem as solving the differentiable optimal transport problem, and then constructing the RNN.

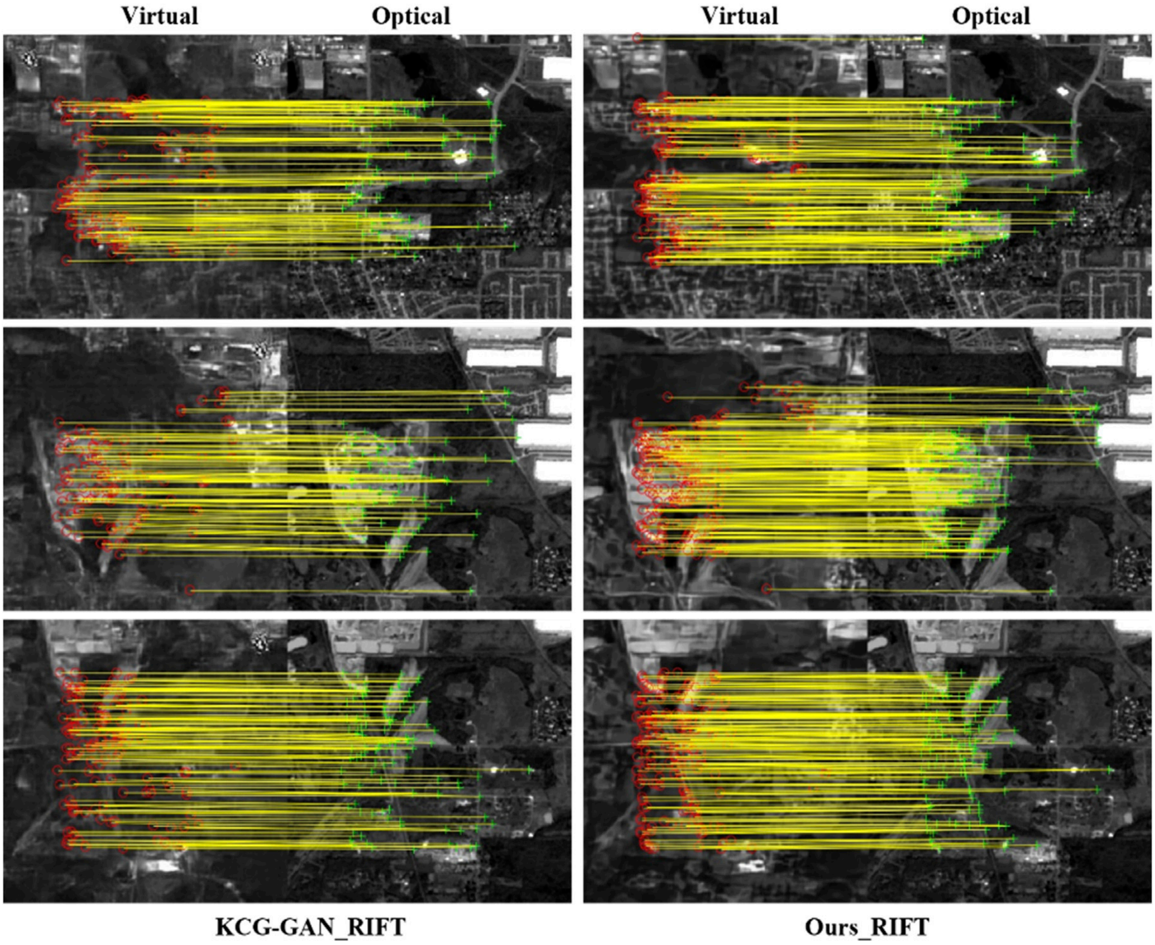


FIGURE 7
Comparison between KCG-GAN and SVMNet with RIFT matching method. The left column of each figure uses KCG-GAN, and the right side is our SVGNet in this paper. On the left side of each set of images are the generated images and on the right side are the optical images.

4.4.1 Visual performance

The traditional feature matching method, RIFT, is selected for feature extraction. We compare the generated networks between KCG-GAN and SVGNet in this paper. Compared with KCG-GAN, our SVGNet virtual maps are more realistic and have high optical consistency. In the texture of KCG-GAN maps, details and surrounding areas are more discordant, and the edges and textures are not as clear as our virtual maps.

From Figure 7, it can be observed that the matching performance of the generated images by KCG-GAN is inferior, with fewer matching points. This can be attributed to the fact that KCG-GAN may not fully preserve the semantic information of the original SAR images during the transformation process to optical images. A comparison between the virtual maps and true optical images may reveal differences in terms of object shape, structure, and other aspects, leading to less accurate matching. Moreover, KCG-GAN's training process may be unstable, such as difficulties in achieving a proper balance between the generator and discriminator or issues such as gradient vanishing or exploding. These factors can hinder network convergence, thereby impacting the quality of generated images and the matching effectiveness. By contrast, our approach demonstrates better matching performance with a higher number of matching points and a higher proportion of correct matches between virtual and optical images. To conclude, our SVGNet generated is superior to the KCG-GAN.

Demonstrated by Figure 8, we compare the matching methods, including LoFTR [27], D2-Net and Superglue [23]. The matching results of our virtual maps and optical images are better than those of the original SAR images and optical images. Considering the fact that the virtual maps generated by our SVGNet can compensate for the loss of information that may occur when the optical and SAR images are considered separately, we can provide a better level of visual information, and we can integrate the visual information and feature representation capabilities of the optical and SAR images. The virtual maps we created retain not only the shape and structure information obtained from SAR on the target, but they also retain the advantages of optical maps in terms of color and detail. These images contained many incorrect matching points, and the number of matching points is relatively small between the original SAR images and the optical images. In contrast, the virtual maps we generated match the optical images better, with more matching points, almost 10 times more than the non-generated matching results, which is a huge improvement, and the results are exciting. It shows that the virtual map generated by our generation network works well.

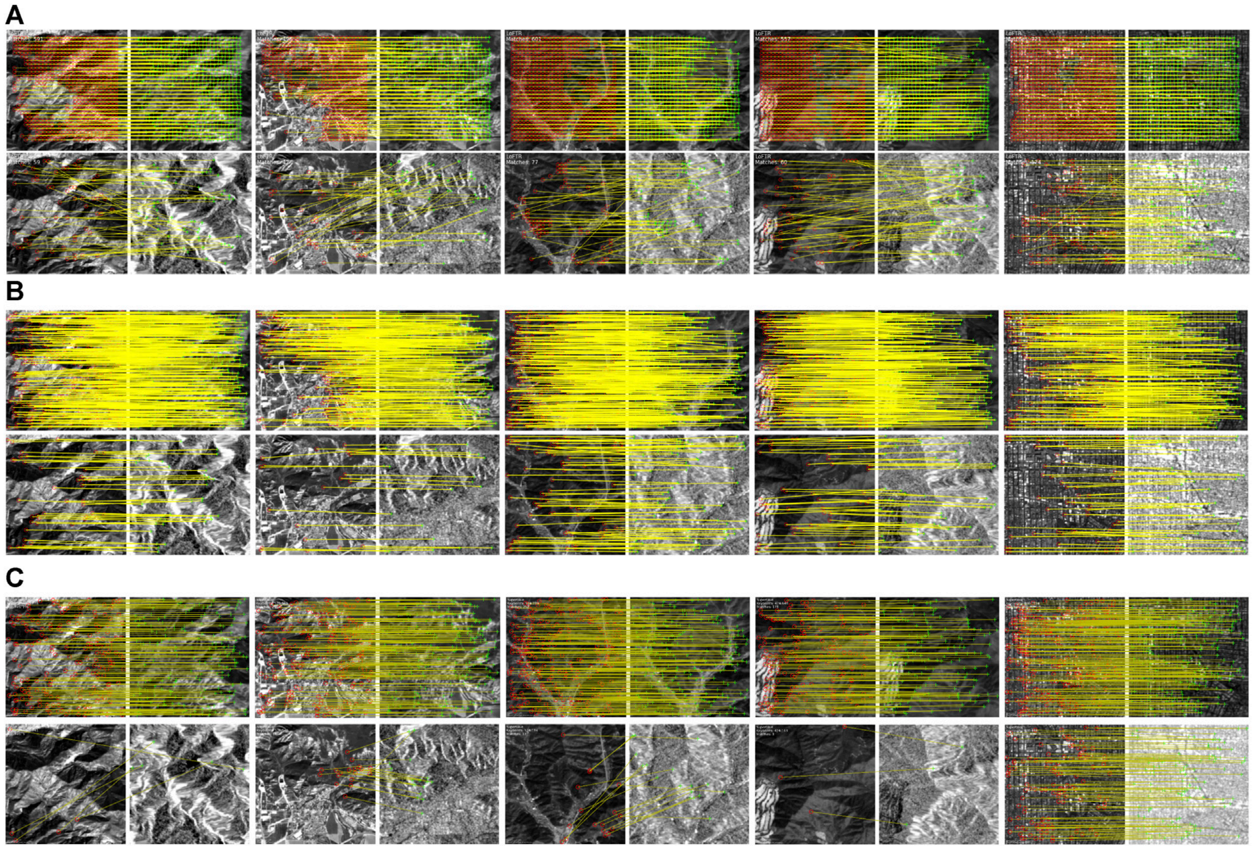


FIGURE 8
Comparison and display of image matching effect before and after generation. **(A–C)** represent three deep learning-based matching methods: **(A)** LoFTR; **(B)** D2-Net; **(C)** Superglue. In each method, the top row showcases the results of image matching after generation, while the bottom row shows the results of image matching before generation.

Overall, SVGNet reduces modal differences and achieves the desired effect. A quantitative analysis of matching methods comparison is presented in the following subsection.

4.4.2 Quantitative analysis

To conduct a quantitative comparison of the effectiveness of our SVGNet, the results are presented in Table 1, which includes a comparison between KCG-GAN and SVGNet, along with a comparison of the generated images before and after applying the three deep learning methods.

The upper part of Table 1 presents comparison of KCG-GAN and SVGNet, showing NCM results and the matching success rate. In three different scenarios, our generative network outperforms KCG-GAN in both NCM and matching success rate. The number of correct matching points is nearly 1.3 times higher than that of KCG-GAN, and our matching success rate (59.78%) is higher than that of KCG-GAN. SVGNet image generation ideas result in a more than double improvement in matching accuracy over direct image matching. Furthermore, our SVGNet improves the RIFT feature matching, indicating the efficiency of the proposed method.

Meanwhile, the bottom half of Table 1 showcases the comprehensive evaluation of three deep learning-based matching methods: LoFTR, D2-Net and Superglue. We use virtual maps generated by the SVGNet to calculate the NCM of matched images and the matching success rate. Prior to the generation of virtual maps, the NCM and matching success rates of the three matching methodologies in the three scenarios were significantly lower. The NCM of LoFTR with the greatest matching effect is almost 85.76 times that of SAR in virtual maps in rural scenes, and 686.05 in urban scenes. In addition, the overall matching success rate of virtual maps using LoFTR matching method reached 95.72%, which was about 4.75 times before the generation. The NCM of D2-Net matching method is about 3.72 times higher after generation, and the matching success rate is also higher than before generation. The NCM of the Superglue matching method in the semi-urban scenario is 27.44 times higher than before, and the matching success rate is also increased by 20.67%. In general, the matching effect after generation has been improved to different degrees under different matching methods. The virtual maps generated by our SVGNet have obtained inspiring results.

Our further evaluation of the accuracy and consistency of image matching consisted of the selection of 20 random images and the manual selection of 10 corresponding checkpoints distributed evenly on the graph after image correction. We use this method

TABLE 1 Quantitative comparison.

Method		NCM			Matching success rate (%)
		Rural	Semi -urban	Urban	
RIFT	KCG-GAN	76.50	79.58	87.77	28.00
	Ours	**123.27**	**132.32**	**143.99**	**59.78**
LoFTR	Optical_SAR	6.40	16.85	16.10	20.15
	Optical_Virtual	**548.90**	**530.10**	**686.05**	**95.72**
D2-Net	Optical_SAR	6.53	5.20	5.90	34.10
	Optical_Virtual	**24.30**	**22.40**	**19.50**	**35.33**
Superglue	Optical_SAR	3.80	3.33	7.58	32.27
	Optical_Virtual	**65.75**	**91.40**	**117.90**	**52.94**

Note that the values in bold are the highest.

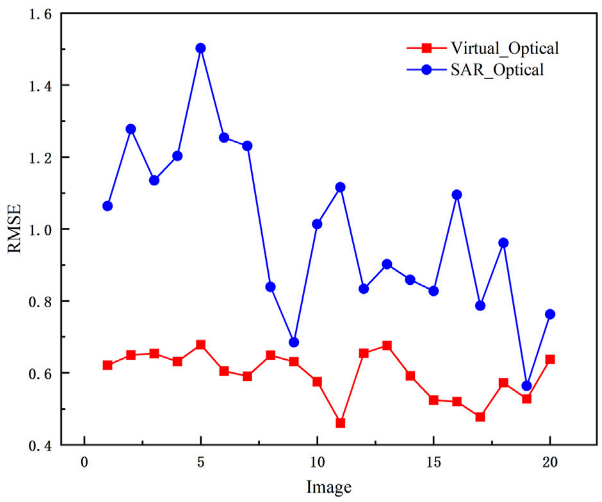


FIGURE 9
Plot of the calculation results for the 20 images used to calculate RMSE.

to determine the degree of difference between the predicted value and the true value, and a smaller RMSE indicates a more reliable prediction. In Figure 9, the virtual map generated by our SVGNet is shown to have lower RMSE than the original SAR image, achieving the lowest RMSE of 0.460 and the highest RMSE of 0.678. Each of the calculated images has a lower RMSE than the original SAR image. Raw SAR images and optical images have a RMSE of 0.564 in the lowest case and 1.502 in the highest case. Consequently, this result indicates that the proposed methodology can enhance matching effectiveness and effectively reduce the noise in the SAR images.

The analysis presented above illustrates the efficacy of the SVGNet for matching images. The evaluation of image matching algorithms using NCM, matching success rate, and RMSE metrics provides comprehensive insights into their performance. As a result of our study, our proposed SVGNet-based method provides superior performance in the generation of virtual maps and in the improvement of image matching accuracy.

4.5 Ablation experiment

In order to evaluate the effectiveness of AG (Attention Gate), SSIM (Structural Similarity), and the kw (sliding window), we conduct a large number of ablation experiments. The following table shows the results of the experiment. Specific experiments are as follows: (a) We remove the AG module from the generator; (b) Instead of using SSIM loss function, L1 is used instead; (c) We modify the size of the sliding window in the discriminator and replace the original 4 with 3, 5 and 7 respectively for the experiment.

The quantitative indicators are summarized in Table 2 below. We select a deep learning matching method LoFTR to evaluate the matching effect of the generated network. From the two indicators shown, removal of AG module, replacement of SSIM and different sliding window sizes will reduce the matching effect. In general, whether it is removing the AG module or replacing the SSIM used, or modifying the size of the sliding window, the matching effect will be reduced. Among them, in the network with AG module removed, although NCM is slightly higher than other methods, it has a certain advantage in matching points. However, to accurately compare the matching accuracy, it is necessary to calculate the MSR (Matching Success Rate). From the results, our results show that it is better than the network without AG module and other networks.

The visual performance of the ablation experiment is as follows. As shown in the Figure 10, in the process of matching images generated by various network modules (image A-E), there is a significant augmentation in the number of visual matching points. Notably, our SVGNet (image F) produces virtual graphs that exhibit superior matching results, characterized by the highest density of corresponding points. This underscores the effectiveness of SVGNet in enhancing the quality and richness of image matching outcomes compared to other modules. In general, the image generated by our SVGNet is better for matching, and the effect is good for different scenes.

TABLE 2 Results of ablation experiments. (Red and blue bold letters represent the optimal and sub-optimal values, respectively. ✗ means not used, ✓ means used and numbers or specific content means alternative content.)

Matching method	Image data	Generative adversarial network			Metrics			
LoFTR	Optical_Virtual	Generator		Discriminator	NCM			MSR
		AG	Loss (SSIM)	kw (4)	Rural	Semi-Urban	Urban	
		✗	✓	✓	**160.99**	**190.30**	156.97	57.16%
		✓	L1	✓	148.98	156.06	214.10	56.36%
		✓	✓	3	158.43	176.97	248.99	**59.16%**
		✓	✓	5	150.71	169.24	244.39	57.69%
		✓	✓	7	152.16	166.44	**249.49**	57.95%
		✓	✓	✓	**160.83**	**185.74**	**252.57**	**59.75%**

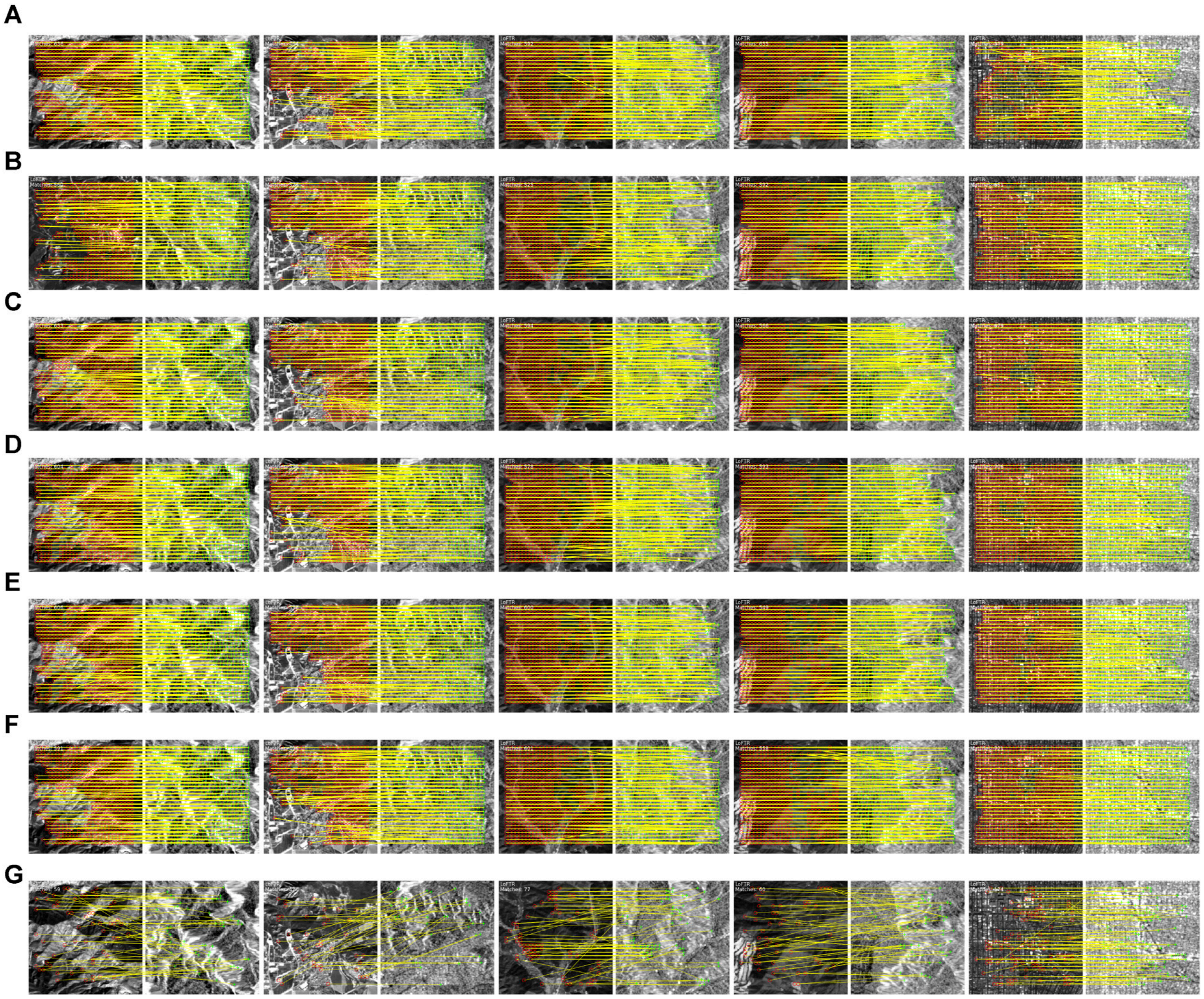


FIGURE 10
Visual performance of the ablation experiment. Use LoFTR for matching. A to H respectively represent: **(A)** no AG module; **(B)** The loss function uses L1 instead of SSIM; **(C)** kw = 3; **(D)** kw = 5; **(E)** kw = 7; **(F)** SVGNet; **(G)** not generated before the match.

5 Conclusion

The paper proposes the Structure Similarity Virtual Map Generation Network as a new generative adversarial network for matching optical and SAR images. The consistency transformation network constructs the U-Net network into a generating network to learn image textures and discover correlations between images. In order to deal with high frequency components effectively and reduce computation, the SSIM is used to reconstruct spatial information to improve image quality. In addition, LSGAN stabilizes GAN training. It has been shown by numerous experiments that NCM and matching success rates are higher for both the comparison network and the comparison before and after the generation, particularly in the more advanced matching method LoFTR, which has an overall matching success rate of 95.72% and a lower RMSE than the non-generated matching method. By using SVGNet in this paper, the virtual maps generated are more realistic. This diminishes the modal difference between SAR and optical images, mitigates the challenge of matching heterosource images and enhances the robustness of the model.

In the future, geometric feature-based approaches can be used to reduce modality differences and improve image alignment in SAR and optical image matching. By incorporating geometric cues and constraints, we aim to achieve more accurate and robust image matching results. This novel perspective will complement existing style transfer-based methods and pave the way for a comprehensive and effective framework for multi-modal image registration and analysis in diverse real-world applications.

Data availability statement

Original datasets are available in a publicly accessible repository: The original contributions presented in the study are publicly available. This data can be found here: https://mediatum.ub.tum.de/1474000.

Author contributions

SC: Conceptualization, Funding acquisition, Project administration, Supervision, Writing–review and editing. LM: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing–original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research is funded by the National Natural Science Foundation of China (62276274); Hubei Province Young Science and Technology Talent Morning Hight Lift Project (202319).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhang L, Zhang L. Artificial intelligence for remote sensing data analysis: a review of challenges and opportunities. *IEEE Geosci Remote Sensing Mag* (2022) 10:270–94. doi:10.1109/mgrs.2022.3145854

2. Yao Y, Zhang Y, Wan Y, Liu X, Yan X, Li J. Multi-modal remote sensing image matching considering Co-occurrence filter. *IEEE Trans Image Process* (2022) 31: 2584–97. doi:10.1109/TIP.2022.3157450

3. Liu J, Zhang Y, Li F. Infrared and visible image fusion with edge detail implantation. *Front Phys* (2023) 11:1180100. doi:10.3389/fphy.2023.1180100

4. Quan D, Wei H, Wang S, Lei R, Duan B, Li Y, et al. Self-distillation feature learning network for optical and SAR image registration. *IEEE Trans Geosci Remote Sensing* (2022) 60:1–18. doi:10.1109/tgrs.2022.3173476

5. Ye Y, Yang C, Zhang J, Fan J, Feng R, Qin Y. Optical-to-SAR image matching using multiscale masked structure features. *IEEE Geosci Remote Sensing Lett* (2022) 19:1–5. doi:10.1109/lgrs.2022.3171265

6. Zhu B, Zhou L, Pu S, Fan J, Ye Y. Advances and challenges in multimodal remote sensing image registration. *IEEE J Miniaturization Air Space Syst* (2023) 4:165–74. doi:10.1109/jmass.2023.3244848

7. Misra I, Rohil MK, Manthira Moorthi S, Dhar D. Feature based remote sensing image registration techniques: a comprehensive and comparative review. *Int J Remote Sensing* (2022) 43:4477–516. doi:10.1080/01431161.2022.2114112

8. Bansal M, Kumar M, Kumar M. 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. *Multimedia Tools Appl* (2021) 80:18839–57. doi:10.1007/s11042-021-10646-0

9. Hassanin A-AIM, Abd El-Samie FE, El Banby GM. A real-time approach for automatic defect detection from PCBs based on SURF features and morphological operations. *Multimedia Tools Appl* (2019) 78:34437–57. doi:10.1007/s11042-019-08097-9

10. Li Z, Zhang H, Huang Y. A rotation-invariant optical and SAR image registration algorithm based on deep and Gaussian features. *Remote Sensing* (2021) 13:2628. doi:10.3390/rs13132628

11. Wang Z, Yu A, Zhang B, Dong Z, Chen X. A fast registration method for optical and SAR images based on SRAWG feature description. *Remote Sensing* (2022) 14:5060. doi:10.3390/rs14195060

12. Jing Y, Yang Y, Feng Z, Ye J, Yu Y, Song M. Neural style transfer: a review. *IEEE Trans visualization Comput graphics* (2019) 26:3365–85. doi:10.1109/TVCG.2019.2921336

13. Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Lett* (2021) 141:61–7. doi:10.1016/j.patrec.2020.07.042

14. Li X, Yu L, Chen H, Fu C-W, Xing L, Heng P-A. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Trans Neural Networks Learn Syst* (2021) 32:523–34. doi:10.1109/tnnls.2020.2995319

15. Abu-Srhan A, Abushariah MAM, Al-Kadi OS. The effect of loss function on conditional generative adversarial networks. *J King Saud Univ - Comput Inf Sci* (2022) 34:6977–88. doi:10.1016/j.jksuci.2022.02.018

16. Ma J, Jiang X, Fan A, Jiang J, Yan J. Image matching from handcrafted to deep features: a survey. *Int J Comput Vis* (2020) 129:23–79. doi:10.1007/s11263-020-01359-2

17. Yang Z, Dan T, Yang Y. Multi-temporal remote sensing image registration using deep convolutional features. *IEEE Access* (2018) 6:38544–55. doi:10.1109/access.2018.2853100

18. Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, et al. *D2-net: a trainable cnn for joint detection and description of local features* (2019). arXiv preprint arXiv:1905.03561.

19. Al-Masni MA, Kim D-H. CMM-Net: contextual multi-scale multi-level network for efficient biomedical image segmentation. *Scientific Rep* (2021) 11:10191. doi:10.1038/s41598-021-89686-3

20. Hao L, Shen P, Pan Z, Xu Y. Multi-level semantic information guided image generation for few-shot steel surface defect classification. *Front Phys* (2023) 11:1208781. doi:10.3389/fphy.2023.1208781

21. Ma W, Zhang J, Wu Y, Jiao L, Zhu H, Zhao W. A novel two-step registration method for remote sensing images based on deep and local features. *IEEE Trans Geosci Remote Sensing* (2019) 57:4834–43. doi:10.1109/tgrs.2019.2893310

22. Zhang H, Ni W, Yan W, Xiang D, Wu J, Yang X, et al. Registration of multimodal remote sensing image based on deep fully convolutional neural network. *IEEE J Selected Top Appl Earth Observations Remote Sensing* (2019) 12:3028–42. doi:10.1109/jstars.2019.2916560

23. Sarlin P-E, DeTone D, Malisiewicz T, Rabinovich A. Superglue: learning feature matching with graph neural networks. *Proc IEEE/CVF Conf Comput Vis pattern recognition* (2020) 4938–47.

24. Ma J, Jiang X, Jiang J, Zhao J, Guo X. LMR: learning a two-class classifier for mismatch removal. *IEEE Trans Image Process* (2019) 28:4045–59. doi:10.1109/tip.2019.2906490

25. Hughes LH, Marcos D, Lobry S, Tuia D, Schmitt M. A deep learning framework for matching of SAR and optical imagery. *ISPRS J Photogrammetry Remote Sensing* (2020) 169:166–79. doi:10.1016/j.isprsjprs.2020.09.012

26. Du W-L, Zhou Y, Zhao J, Tian X. K-means clustering guided generative adversarial networks for SAR-optical image matching. *IEEE Access* (2020) 8: 217554–72. doi:10.1109/access.2020.3042213

27. Sun J, Shen Z, Wang Y, Bao H, Zhou X. LoFTR: detector-free local feature matching with transformers. *Proc IEEE/CVF Conf Comput Vis pattern recognition* (2021) 8922–31.

28. Zhang H, Le Z, Shao Z, Xu H, Ma J. MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Inf Fusion* (2021) 66:40–53. doi:10.1016/j.inffus.2020.08.022

29. John D, Zhang C. An attention-based U-Net for detecting deforestation within satellite sensor imagery. *Int J Appl Earth Observation Geoinformation* (2022) 107: 102685. doi:10.1016/j.jag.2022.102685

30. Kumar MS, Ganesh D, Turukmane AV, Batta U, Sayyadliyakat KK. Deep convolution neural network based solution for detecting plant diseases. *J Pharm Negative Results* (2022) 464–71. doi:10.47750/pnr.2022.13.S01.57

31. Lutfhi A, Rumini B. The effect of layer batch normalization and droupout of CNN model performance on facial expression classification. *JOIV: Int J Inform Visualization* (2022) 6:481–8. doi:10.30630/joiv.6.2-2.921

32. Macêdo D, Zanchettin C, Oliveira ALI, Ludermir T. Enhancing batch normalized convolutional networks using displaced rectifier linear units: a systematic comparative study. *Expert Syst Appl* (2019) 124:271–81. doi:10.1016/j.eswa.2019.01.066

33. Li J, Su J, Xia C, Ma M, Tian Y. Salient object detection with purificatory mechanism and structural similarity loss. *IEEE Trans Image Process* (2021) 30:6855–68. doi:10.1109/TIP.2021.3099405

34. Lee C-K, Cheon Y-J, Hwang W-Y. Least squares generative adversarial networks-based anomaly detection. *IEEE Access* (2022) 10:26920–30. doi:10.1109/access.2022.3158343

35. Schmitt M, Hughes LH, Qiu C, Zhu XX. *SEN12MS–A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion* (2019). arXiv preprint arXiv:1906.07789.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 14 February 2024
DOI 10.3389/fphy.2024.1364372

EDITED BY
Bo Xiao,
Imperial College London, United Kingdom

REVIEWED BY
Huafeng Li,
Kunming University of Science and Technology, China
Puhong Duan,
Hunan University, China
Youlin Wang,
University of Montreal, Canada

*CORRESPONDENCE
Dongli Fan,
fdltmmu@sina.com
Zeyuan Lei,
leizeyuan0854@163.com

RECEIVED 02 January 2024
ACCEPTED 02 February 2024
PUBLISHED 14 February 2024

CITATION
Xiong Y, Yu K, Lan Y, Lei Z and Fan D (2024), Hair cluster detection model based on dermoscopic images.
Front. Phys. 12:1364372.
doi: 10.3389/fphy.2024.1364372

Hair cluster detection model based on dermoscopic images

Ya Xiong[1], Kun Yu[2], Yujie Lan[1], Zeyuan Lei[1]* and Dongli Fan[1]*

[1]Department of Plastic and Cosmetic Surgery, Xinqiao Hospital, The Army Medical University, Chongqing, China, [2]College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China

Introduction: Hair loss has always bothered many people, with numerous individuals potentially facing the issue of sparse hair.

Methods: Due to a scarcity of accurate research on detecting sparse hair, this paper proposes a sparse hair cluster detection model based on improved object detection neural network and medical images of sparse hair under dermatoscope to optimize the evaluation of treatment outcomes for hair loss patients. A new Multi-Level Feature Fusion Module is designed to extract and fuse features at different levels. Additionally, a new Channel-Space Dual Attention Module is proposed to consider both channel and spatial dimensions simultaneously, thereby further enhancing the model's representational capacity and the precision of sparse hair cluster detection.

Results: After testing on self-annotated data, the proposed method is proven capable of accurately identifying and counting sparse hair clusters, surpassing existing methods in terms of accuracy and efficiency.

Discussion: Therefore, it can work as an effective tool for early detection and treatment of sparse hair, and offer greater convenience for medical professionals in diagnosis and treatment.

KEYWORDS

hair loss, dermatoscope, hair cluster detection, feature fusion, dual attention module

1 Introduction

As a common issue, sparse hair [1] brothers many people, affecting both males and females alike [2], [3]. Hair loss or thinning primarily attributed to genetic factors, hormonal changes, environmental conditions, or medical conditions is a prevalent problem affecting millions worldwide [4]. Regardless of gender or age, it impacts an individual's self-esteem, personal aesthetics, and overall mental health. Traditional solutions such as drug treatments, hair transplants, or wearing wigs have achieved varying degrees of success and affordability, but they do not fundamentally resolve the problem or prevent its recurrence. Therefore, early detection and predictive analysis of sparse hair conditions are vital for implementing preventative measures and more effective treatments [5].

Over the past few decades, both domestic and international researchers have been exploring how to accurately detect sparse hair. The earliest research primarily relies on manual feature extraction and traditional image processing techniques [6]. However, due to the limitations on the selection and representational power of features, these methods are difficult to adapt to the complex and diverse forms of hair clusters. Therefore, with the rapid development of computer vision and deep learning [7], researchers introduce neural network into the field of sparse hair target detection. In recent years, with the advent of artificial intelligence (AI) and deep learning technologies, their application in the healthcare sector grows exponentially, providing promising results in different fields like diagnosis, prognosis, treatment planning, and public health [8]. In light of this, the development of AI-driven sparse hair detection models [9], especially those based on neural network, offers a promising research pathway.

Based on the strong learning capability and adaptability, neural network is able to learn effective feature representations from a large amount of data and train and optimize through the backpropagation algorithm. This provides new opportunities and challenges for the target detection of sparse hair [10]. Researchers design and improve hair cluster target detection models based on neural network to enhance detection accuracy and robustness.

At present, domestic and international research in the field of sparse hair detection is still in the exploratory stage [11]. Some studies have utilized traditional Convolutional Neural Network (CNN) to detect hair clusters, improving detection performance by constructing deep-level feature representations and using effective loss functions. Other studies have explored more advanced network structures, such as Recurrent Neural Network (RNN) and Attention Mechanisms, to capture the temporal information and local details of hair clusters. In summary, using neural network in hair cluster target detection models for sparse hair detection has enormous potential to thoroughly transform hair care and treatment [12].

However, the target detection of sparse hair still faces some challenges. Hair clusters exhibit diverse morphologies with differences in color, texture, and shape [13], posing difficulties for detection algorithms. Additionally, due to the sparse distribution of hair, hair cluster targets unevenly occupy proportions in images, making target detection more challenging. Currently, dermatoscopy is a non-invasive diagnostic technique that allows the observation of hair shafts, follicles, and capillaries, providing a visual representation of inflammation around the scalp and changes in hair shaft diameter and shape [14]. It is widely used in the diagnosis and treatment of hair diseases, as well as in the assessment and follow-up of prognosis [15], [16], [17], [18]. Digital intelligent analysis of dermatoscopy is still in the developmental stage, and research on dermoscopy for androgenetic alopecia is limited. For the daily management and assessment of treatment outcomes for patients with hair loss, hair counting plays a crucial role. However, there are currently no clear standards for a comprehensive evaluation of hair loss across the entire scalp.

In response to these challenges, this study utilizes hair images obtained by dermoscopy, combined with existing advanced target detection techniques, to propose an efficient and accurate sparse hair cluster target detection model. This model sets the hair cluster as the detection target (in this paper, the sparse hair or hair loss area) and predicts the number of hair clusters. This paper has three main contributions as follows.

1. Based on the advanced existing object detection networks, a dermoscopy image hair detection network structure based on an improved object detection neural network is proposed to better adapt to sparse hair detection. Through experiments, it proves that the proposed method surpasses the existing methods in terms of accuracy and efficiency, providing an effective tool for early detection and treatment of sparse hair.
2. Multi-Level Feature Fusion Module: A new multi-level feature fusion Module (MLFF) is designed to extract and fuse features at different levels. The MLFF structure can obtain features from different convolutional layers, then integrate these features through a specific fusion strategy to produce a richer, more representative feature expression.
3. Channel-Space Dual Attention Module: A new attention mechanism, the Channel-Space Dual Attention Module, is proposed to consider both channel and spatial dimensions' information simultaneously. The CSDA module can handle channel and spatial correlation in a unified framework, thereby further enhancing the model's expressive capacity and accuracy of sparse hair detection.

2 Related work

With the rapid development of computer technology and computer-assisted medical diagnostic systems, the continuous growth of computational power and data, deep learning has experienced tremendous development, becoming one of the powerful tools in the medical field. The technology of feature extraction and classification from medical images [19], [20] using maturing deep learning models is increasingly mature.

The field of object detection has always been a research hotspot. For instance, one study proposed a safety helmet detection method based on the YOLOv5 algorithm [21]. This research involved annotating a collected dataset of 6,045, training, and testing the YOLOv5 model with different parameters. In another study, YOLOv4 was employed for small object detection and anti-complex background interference in remote sensing images [22]. With the use of deep learning-based algorithms, ship detection technology has greatly enriched, allowing monitoring of large, distant seas. Through the use of a custom dataset with four types of ship targets, Kmeans++ clustering algorithm for prior box framework selection, and transfer learning method, the study enhanced YOLOv4's detection ability. Further improvements were introduced by replacing Spatial Pyramid Pooling (SPP) with a Receptive Field Block with dilated convolution and adding a Convolutional Block Attention Module (CBAM). These modifications have improved the detection performance of small vessels and enhanced the model's resistance to complex backgrounds. Due to the relatively large size and distinct features of vessels, the detection results are satisfactory. However, it remains a challenge for densely packed, small targets.

In recent years, there has been an emergence of research utilizing deep learning methods in skin imaging analysis, particularly in studies related to hair. Researchers have explored the application of deep learning-based object detection [23], [24], segmentation [25], and other algorithms in hair detection and segmentation. These studies primarily focus on aspects such as hair detection, removal, segmentation, and even reconstruction, but there is room for improvement in terms of accuracy.

Various deep learning structures and techniques are introduced in multiple studies to address the challenges related to hair recognition and removal in dermoscopic images. One such

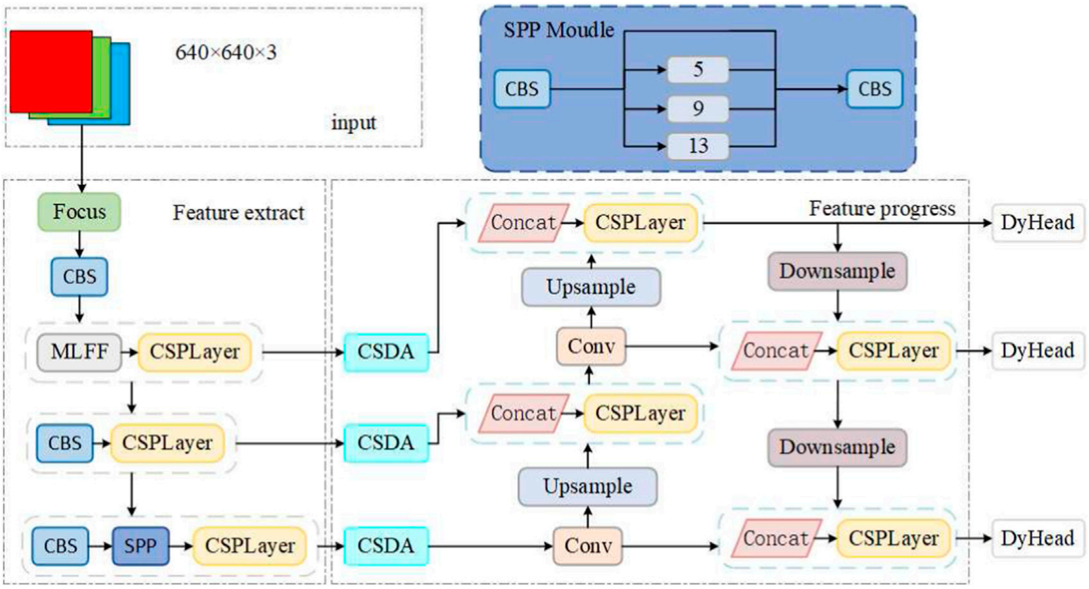


FIGURE 1
The method proposed in this paper.

study proposed a novel deep learning technique, Chimera Net [26], an encoder-decoder architecture that uses a pretrained EfficientNet and squeeze-and-excitation residual (SERes) structure. This method exhibited superior performance over well-known deep learning methods like U-Net and ResUNet-a. Additionally, other research explored difficulties and solutions related to hair reconstruction. A novel method was proposed to capture high-fidelity hair geometry with strand-level accuracy [13]. The multi-stage approach includes a new multiview stereo method and a novel cost function for reconstructing each hair pixel into a 3D line. The task of Digital Hair Removal (DHR) also received ample research. One study proposed a DHR deep learning method using U-Net and free-form image restoration architecture [9]. It outperforms other state-of-the-art methods on the ISIC2018 dataset. Another study explored a similar theme Attia et al. [10], highlighting the challenges associated with hair segmentation and its impact on subsequent skin lesion diagnosis. Moreover, one paper delved into an important metric for determining the number of hairs on the scalp [27]. It stressed the need for an automated method to increase speed and throughput while lowering the cost of counting and measuring hair in trichogram images. The proposed deep learning-based, enables rapid, fully automatic hair counting and length measurement. Another study described a real-time hair segmentation method based on a fully convolutional network, the basic structure of which is an encoder-decoder [28]. This method uses Mobile-Unet, a variant of the U-Net segmentation model, which combines the optimization techniques of MobileNetV2.

In summary, the above studies emphasize the enormous potential of deep learning techniques in advancing hair-related dermoscopy research. However, deep learning-based sparse hair detection is still in the exploratory stage. To address these challenges, this paper, based on sparse hair dermoscopic medical images, proposes a dermoscopic image hair detection network structure based on an improved object detection neural network to achieve the detection of sparse hair clusters (sparse hair or hair loss areas in this paper) and predict the number of hair clusters.

3 Materials and methods

In this section, we will provide a detailed introduction to the proposed sparse hair detection network structure, which is based on the object detection network [29]. Firstly, we will describe the overall structure of the network in Section 3.1. Subsequently, we will highlight the novel contributions of this paper in Sections 3.2, 3.3, namely, the MLFF Module and the CSDA Module, respectively.

3.1 Overall structure

The overall framework proposed for sparse hair detection in this article is illustrated in Figure 1, primarily based on enhancements to classical object detection network architectures. Given the crucial significance of the accuracy of the sparse hair detection model for hair target recognition and assisting doctors in obtaining diagnostic results, the model proposed in this article is intended for application in sparse hair target detection models.

It can be divided into three parts: the feature extraction backbone network, the feature enhancement and processing network, and the detection network. Specifically, the feature extraction backbone network is a convolutional neural network that incorporates the concept of a feature pyramid architecture, capable of extracting image features at different levels and reducing model computation while speeding up training. As shallow features contain more semantic information, a MLFF Module is proposed to handle them, preventing the loss of semantic information. At the end of the feature extraction backbone network, there is a Spatial Pyramid Pooling (SPP) module aimed at improving the network's receptive field by transforming feature maps of arbitrary sizes into fixed-size feature vectors. Three main backbone features can be obtained through the feature extraction backbone network.

In the feature enhancement and processing network, the Channel-Spatial Dual Attention module (CSDA) is introduced. The three feature layers obtained from the backbone network undergo processing through this module to generate enhanced features. Subsequently, processing is carried out based on the YOLOv5 network model. This network segment primarily consists of a series of feature aggregation layers that mix and combine image features to generate a Feature Pyramid Network (FPN). The output feature maps are then transferred to the detection network. With the adoption of a novel FPN structure, this design strengthens the bottom-up pathway, improving the transfer of low-level features and enhancing the detection of objects at different scales. Consequently, it enables the accurate identification of the same target object with varying sizes and proportions.

The detection network is primarily employed for the final detection phase of the model. It applies anchor boxes to the feature maps output from the preceding layer and outputs a vector containing the class probability, object score, and position of the bounding box around the object. The detection network of the proposed architecture consists of three detection layers, with inputs being feature maps of sizes 80 × 80, 40 × 40, and 20 × 20, respectively, used for detecting objects of different sizes in the image. Each detection layer ultimately outputs an 18-dimensional vector ((4 + 1+1)×3 anchor boxes). The first four parameters are used for determining the regression parameters for each feature point, and adjusting these regression parameters yields the predicted box. The fifth parameter is utilized to determine whether each feature point contains an object, and the last parameter is employed to identify the category of the object contained in each feature point. Subsequently, the predicted bounding boxes and categories of the targets in the original image are generated and labeled, enabling the detection of clusters of hair targets in the image.

Algorithm 1 describes the training process of the hair detection model in dermoscopic images. The computation time increases linearly with the increase of training sample, batch size, and training epochs. The time complexity of the training algorithm is $O\ [E \times (n/B) \times 2 \times (M - 1)]$.

```
Input: Training dataset $D$, segmentation model $M$, number of epochs $E$, learning rate $\eta$, $n$ training samples, loss function $L$, batch size $B$
Output: Trained segmentation model $\hat{M}$.
Initialize segmentation model $M$
for $e \in [1, E]$ do
    for $b \in [1, n/B]$ (mini-batch $b$ in $D$ with size $B$) do
        Perform forward pass on $M$ with mini-batch $b$
        Calculate detection loss according to the loss function $L$
        Perform backward pass and update model weights and model according to the gradient
    end for
    Save the trained model $\hat{M}$
end for
```

Algorithm 1. A dermoscopy-image hair detection model based on improved object detection neural network.

3.2 Multi-level feature fusion structure

The main task of the MLFF (Multi-Level Feature Fusion) structure is to process a large amount of semantic information contained in shallow layers. Its structure is shown in Figure 2. The purpose of this module is to extract and fuse semantic information from shallow features, so that the resulting feature information is more detailed and more suitable for subsequent object detection tasks. Semantic feature information reflects a global feature of homogeneous phenomena in the image, depicting the surface organization and arrangement rules of slow-changing or cyclically-changing structures in the image. However, the low-level information extracted by the original backbone network (such as pixel values or local region attributes) is often of low quality and contrast, making it difficult to obtain and utilize this low-level information effectively. This paper proposes the MLFF module to address this problem.

As shown in Figure 2, in this module, a feature X_1 Eq. 1 before the output of this module serves as the input. It undergoes two consecutive CBS modules, resulting in two feature layers X_2 and X_3 Eq. 1, represented as follows:

$$\begin{aligned} X_1 &\in \mathbb{R}^{H\times W\times C} \\ X_2 &\in \mathbb{R}^{H\times W\times C} \\ X_3 &\in \mathbb{R}^{H\times W\times C} \end{aligned} \tag{1}$$

The CBS module represents a sequence of convolution operation, batch normalization operation, and activation function operation. This sequence is designed to capture local relationships within the input data, facilitating effective feature learning in images. Simultaneously, it helps mitigate the vanishing gradient problem and enhances the model's adaptability to changes in the distribution of input data. The CBS module can be expressed as follows:

$$X_{out} = SiLU\{BN\,[Conv\,(X_{in}, c_{in}, c_{out})]\} \tag{2}$$

Where Conv represents the convolution operation, BN represents batch normalization operation, and SiLU represents the activation function operation. X_{out} represents the output feature of the CBS module, X_{in} represents the input feature of the CBS module, c_{in} represents the number of channels in the input feature, and c_{out}represents the number of channels in the output feature.

After the three features obtained through stacking and fusion, two feature layers are obtained. They will undergo another CBS module (where $c_{in} = c_{out}$) for feature processing. Finally, these features will be stacked together, achieving feature integration. With the depth of feature processing and fusion, the dimension of the image feature vector continuously increases, and the size of each slice changes accordingly. Finally, after passing through a CBS module (where $c_{in} = c_{out}$), as in Eq. 2, the output feature Eq. 3 is:

$$X_{MLFF} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C} \tag{3}$$

The obtained features will be inputted into the feature enhancement and processing network for further processing, where the abundant semantic information contained in the shallow layers will be fully utilized to achieve better detection performance. The first three branches actually correspond to dense residual structures, which take into account the easy-to-

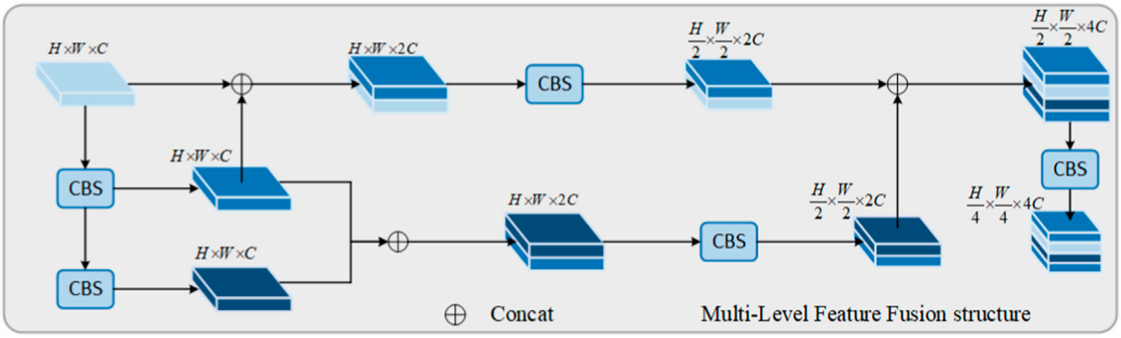

$H \times W \times C$
$H \times W \times 2C$
CBS
$\frac{H}{2} \times \frac{W}{2} \times 2C$
$\frac{H}{2} \times \frac{W}{2} \times 4C$
$\frac{H}{4} \times \frac{W}{4} \times 4C$
Concat
Multi-Level Feature Fusion structure

FIGURE 2
Multi-level feature fusion structure.

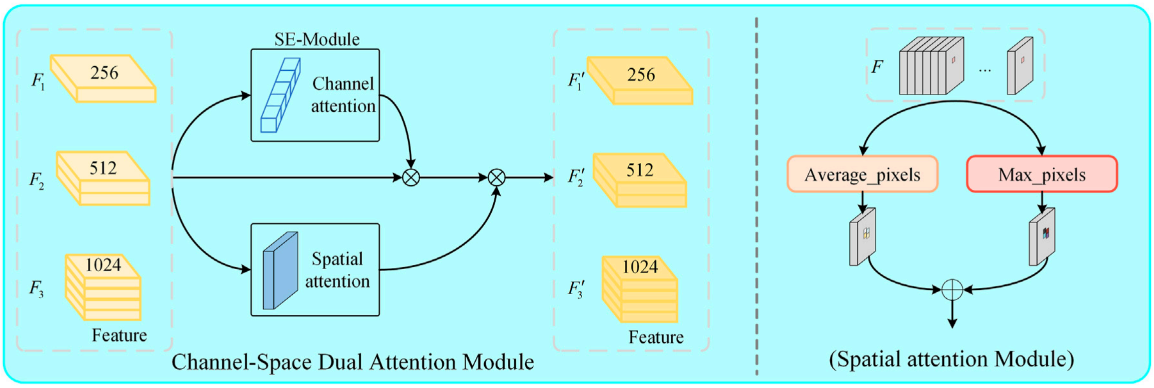


FIGURE 3
Channel spatial dual attention module.

optimize characteristics of residual networks, and the ability of residual networks to improve the overall accuracy of the network by adding a considerable depth. In addition, skip connections are used to alleviate the problem of gradient disappearance caused by the depth of the neural network.

For the CBS module, the SiLU activation function is used, which is an improved version based on the Sigmoid activation function and ReLU activation function. SiLU has the characteristics of no upper bound and a lower bound, smoothness, and non-monotonicity. SiLU performs better than ReLU in deep models and can be considered as a smoothed ReLU activation function. Its specific implementation is shown in the equation below Eq. 4:

$$f(x) = x \cdot sigmoid(x) \tag{4}$$

3.3 Channel-space dual attention module

After obtaining feature information at different depths, it is necessary to further process these features to capture the target information in them. Therefore, this paper proposes a Channel-Space Dual Attention Module (CSDA) for feature inference, as shown in Figure 3. Finally, the inferred information is passed through the second part of the object detection model architecture to obtain three types of feature maps.

The module proposed in this article takes the feature layers obtained from the feature extraction backbone network, namely, $F_1 \in \mathbb{R}^{80\times80\times256}$, $F_2 \in \mathbb{R}^{40\times40\times512}$ and $F_3 \in \mathbb{R}^{20\times20\times1024}$, and infers attention maps along two different dimensions. One dimension is the channel attention mechanism, which is based on the SE module [30] and uses global average pooling to calculate channel attention. The other dimension is the spatial attention mechanism, which focuses on which pixels in different feature maps are important and require significant attention. Then, the channel attention map and the spatial attention map are multiplied successively with the feature maps on the backbone to perform adaptive feature focusing, resulting in corresponding feature maps F'_1, F'_2 and F'_3.

For the Squeeze-and-Excitation module, it can be viewed as a computational unit that mainly embeds the dependency factors of feature map channels into variable v. This is to ensure that the network can enhance its sensitivity to information features and suppress less useful features. In the channel-wise optimization process, squeezing and excitation steps are applied to optimize the response of the convolutional kernel, in order to capture the correlation of channel information. The specific implementation is shown in the following equation:

$$C_{tran}: x \to y;\ x, y \in \mathbb{R}^{H\times W\times C} \tag{5}$$

In the equation, C_{tran} is the convolutional operator, $v = [v_1, v_2, \ldots, v_n]$ represents the learned weights in the network, and n denotes the parameters of the $n - th$ convolutional kernel. Therefore, the output of the convolutional operator is $Y = [y_1, y_2, \ldots, y_n]$, which is implemented as shown in Eq. 5 and Eq. 6. In the proposed attention module, after the channel attention, we can obtain the feature $F_{channel}$.

$$Y = v * X = \sum_{n=1}^{n} v_n * x_n \tag{6}$$

Regarding the spatial attention module, as shown in the right half of Figure 3, the feature map obtained by the feature extraction network is understood as a three-dimensional space, where each slice corresponds to a channel. Firstly, the values at the same position on different channels are subjected to average pooling and max pooling operations to obtain the features $F_{\max}$, $F_{average}$ Eq. 7.

$$\begin{aligned} F_{\max} &= MaxPool(F) \\ F_{average} &= AvgPool(F) \end{aligned} \tag{7}$$

Finally, convolution and normalization operations are applied to generate a 2D spatial attention map $F_{spatial}$, which is computed as follows Eq. 8:

$$F_{spatial} = sigmoid\left(f^{7\times 7}\left(F_{\max}, F_{average}\right)\right) \tag{8}$$

The symbol $f^{7} \times {}^{7}$ represents a convolution operation with a kernel size of 7 × 7. After obtaining the channel attention map, it is multiplied with the input feature map F to obtain a new feature map F'. This new feature map F' is then multiplied with the spatial attention map to obtain the final feature map F''. The overall process can be described as follows Eq. 9:

$$\begin{aligned} F' &= F_{channel} \otimes F \\ F'' &= F' \otimes F_{saptial} \end{aligned} \tag{9}$$

Finally, three feature maps, denoted as F'_1, F'_2 and F'_3, can be obtained. The obtained new features are then processed and enhanced using feature processing networks and detection networks to obtain the final object detection results. The experimental results of the proposed network will be discussed in Section 3 of this paper.

3.4 Attention dynamic head

Introducing dynamic heads [31], based on three feature maps F'_1, F'_2 and F'_3, the general formula for applying self-attention is as follows Eq. 10:

$$W(\mathcal{F}) = \pi(\mathcal{F}) \cdot \mathcal{F} \tag{10}$$

Where $\pi(\cdot)$ is an attention function. A simple solution to this attention function is achieved through fully connected layers. However, due to the high dimensionality of tensors, directly learning attention functions across all dimensions is computationally expensive and practically unaffordable. Therefore, transforming the attention function into attention along three directions, with each attention focusing on a single direction, is proposed Eq. 11.

$$W'(\mathcal{F}) = \pi_C\left(\pi_S\left(\pi_L(\mathcal{F}) \cdot \mathcal{F}\right) \cdot \mathcal{F}\right) \cdot \mathcal{F} \tag{11}$$

Where $\pi_L(\cdot)$, $\pi_S(\cdot)$, $\pi_C(\cdot)$ are three different attention functions applied respectively to dimensions L, S, and C.

4 Experimental results and analysis

4.1 Datasets

In the experiment described in this paper, both the training and testing datasets are sourced entirely from hospitals and collected based on different patients, each with varying degrees of hair sparsity. The original dataset is devoid of any annotations, and labeling is used to annotate it, generating XML-format files to store the labeled tags. Each image corresponds to one XML file, containing multiple hair cluster labels, primarily annotating each hair cluster. In the experiment, each hair cluster does not exceed three strands. A total of 200 images were annotated for the dataset. As neural network-based object detection models are developed on the basis of extensive image data, the dataset is expanded and divided through data augmentation, resulting in 500 images. From these, 50 images are randomly selected as the validation set, and another 50 images are chosen as the test set. This is done to enrich the dataset size, better extract features of hair belonging to different labeled categories, and prevent the trained model from overfitting. The objective of this dataset is to achieve hair detection in populations with sparse hair, identifying the number of hair clusters.

4.2 Experimental details

During the preprocessing stage, the source dataset had a size of 1,920 × 1,080. In this study, all hair datasets underwent image enhancement and partitioning, resulting in a final size of 640 × 640 for each slice.

In the experiment, all programs were implemented in the PyTorch framework under the Windows 10 operating system. The training process used one GeForce RTX 3090 GPU and was written in Python language, calling CUDA, CuDNN, OpenCV, and other required libraries. The optimizer used in the experiment was SGD, with a momentum of 0.937 and default parameters for other settings. The initial learning rate, weight decay, and batch size were set to 0.01, 5e-4, and 8, respectively, and the epoch was set to 500. The trained model's weight file was saved, and the model's performance was evaluated using the test set.

The model evaluation metrics adopted include commonly used object detection metrics such as Precision, Recall, mAP (mean average precision), and F1 score, which are used to assess the performance of the trained model. Visual comparison was also conducted. The implementation of these metrics is as follows Eq. 12:

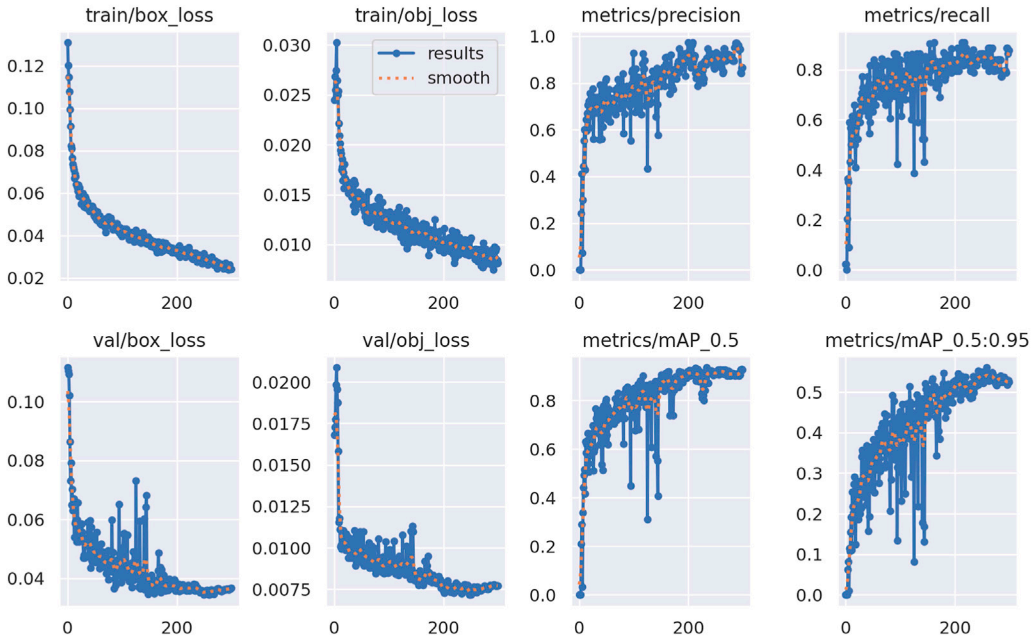


FIGURE 4
Network training situation.

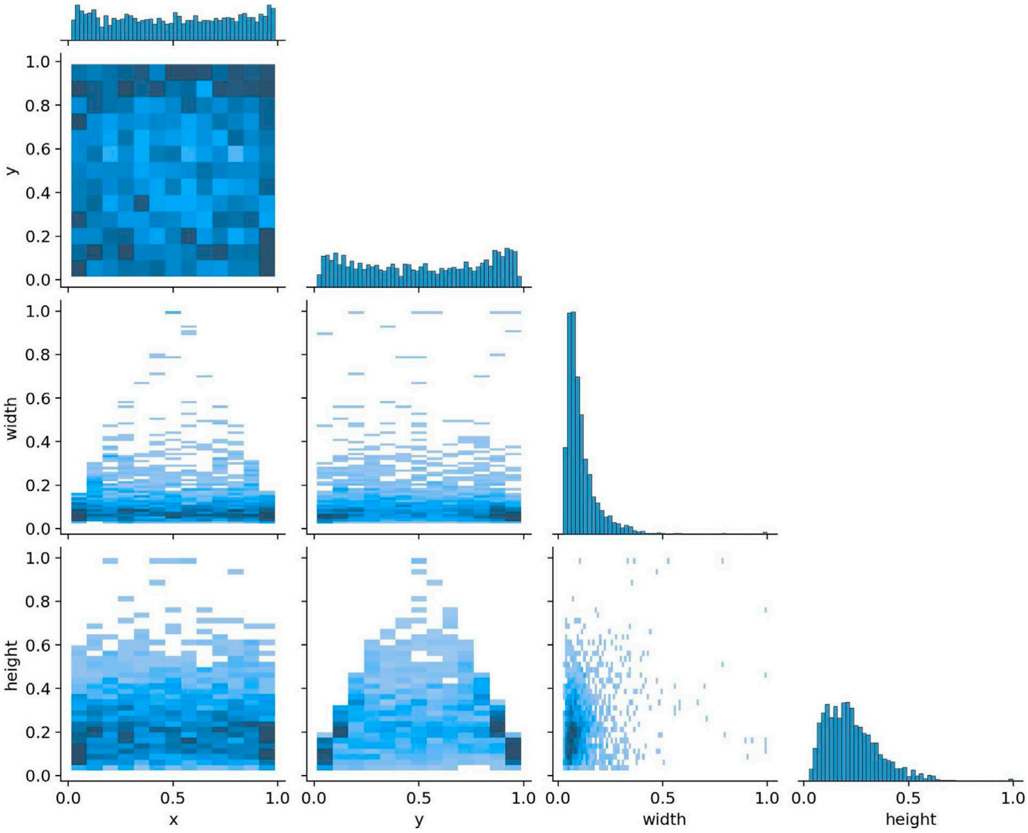


FIGURE 5
Correlation between predicted labels during network training.

TABLE 1 Comparison with different detection networks (Bold numbers represent best results).

Networks	year	Precision	mAP	F1 score	Recall
YOLOv3	2018	0.733	0.500	0.35	0.471
YOLOv4	2020	0.768	0.561	0.58	0.434
Mobilenet YOLOv4	2020	0.792	0.406	0.21	0.245
YOLOv5	2020	0.865	0.706	0.63	0.677
Detr	2020	0.822	0.717	0.65	0.854
FastestV2	2021	0.479	0.458	0.52	0.564
YOLOv7	2022	0.816	0.697	0.66	0.691
FastestDet	2022	0.609	0.524	0.47	0.593
YOLOv8	2023	0.820	0.658	0.63	0.712
Our network	-	**0.898**	**0.734**	**0.72**	**0.873**

TABLE 2 Comparison of ablation experiments of target detection indicators on data sets (Bold numbers represent best results).

Networks	Precision	mAP	F1 score	Recall
Without MLFF	0.817	0.680	0.57	0.712
Without CSDA	0.762	0.599	0.33	0.588
Our network	**0.898**	**0.734**	**0.72**	**0.873**

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ mAP &= \frac{1}{C}\sum_{k=0}^{C} AP_k \\ F1 &= \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \end{aligned} \tag{12}$$

Among them, *TP* represents the number of correctly identified clusters of hair; *FP* represents the number of clusters mistakenly identified as hair; *FN* represents the number of hair cluster targets that were not successfully identified; *C* represents the number of categories of hair cluster targets; *AP* represents the area enclosed by the precision-recall curve and the coordinate axis.

Figure 4 displays the training and validation loss curves, as well as precision, recall, and mAP curves for the entire training process. The model is trained from scratch, and from the curves in the figure, it is evident that the network model descends rapidly in the first 50 epochs and gradually stabilizes thereafter. In the figure, a smaller box_loss indicates more accurate bounding boxes, and a smaller obj_loss indicates more accurate predictions of targets. Precision, recall, and mAP curves stabilize later, indicating a good training outcome. In summary, the figure demonstrates that the model for hair cluster detection is well-trained and does not exhibit overfitting. Figure 5 shows the correlation between predicted labels during the training process of the hair cluster object detection model. Figure 5 is a set of 2D histograms, illustrating the contrast between each axis of the data. Labels in the image are located in the xywh space, where x and y represent the center values of the label box, and w and h represent the length and width of the label box. The histograms of x and y in Figure 5 indicate that the size variation of detected targets is small. Additionally, the distribution plots of x and width, as well as y and height, show that their relationships have a linear correlation. Combined with Figure 4, this suggests that the proposed model for the hair cluster object detection task is trainable.

4.3 Comparative experiments

In the comparative experiments, to validate the performance of the proposed hair cluster detection model based on sparse hair, experiments and analyses were conducted on test set images using publicly available source code of classical object detection models. The object detection network developed in this study was compared with YOLOv3 [32], YOLOv4 [33], MobileNet YOLOv4, YOLOv5, Detr, FastestV2, YOLOv7, FastestDet, and YOLOv8 on test set images. Table 1 presents the performance of the proposed method and other methods on the test set.

The comparative experimental results in Table 1 indicate that the hair cluster detection model proposed in this study achieves the highest mAP value, surpassing the classical YOLOv5 network model by 2.8%. Additionally, it outperforms the latest YOLOv8 by 7.6%. This suggests that the proposed algorithm has advantages in the task of hair cluster target recognition. Moreover, the proposed model achieves the highest Precision, F1, and Recall scores, demonstrating the superior performance of the sparse hair cluster model proposed in this study. Therefore, the results indicate that the proposed model can ensure accurate identification of sparse hair clusters, comparable to the best methods in terms of metrics, and surpassing most other methods.

To more clearly illustrate the performance of the proposed method, visual experiments were conducted on six images selected from the test set, as shown in Figure 6. Figure 6 displays the visual comparison of hair cluster detection results obtained by the proposed method and five other methods (YOLOv8, YOLOv7, Detr, FastestDet, FastestV2) under the same experimental conditions. It is evident that the proposed method achieves more accurate hair cluster detection results compared to other methods.

As evident from the obtained detection results above, the proposed hair cluster detection model for sparse hair in this study has achieved significant results. Simultaneously, the algorithm accomplishes counting and visualizing the detected clusters. A comparison reveals that the method developed in this study exhibits the best performance in hair cluster detection. In Figure 6, it can be observed that other methods show instances of hair cluster omission. In summary, the method investigated in this study demonstrates commendable hair cluster detection performance. Finally, for a more comprehensive comparison of the advantages of the proposed method against different approaches, Figure 7 depicts bar charts representing the hair cluster detection performance of various methods across different metrics. The performance on four metrics is illustrated separately. It is evident that the proposed method holds a significant advantage in hair cluster detection tasks.

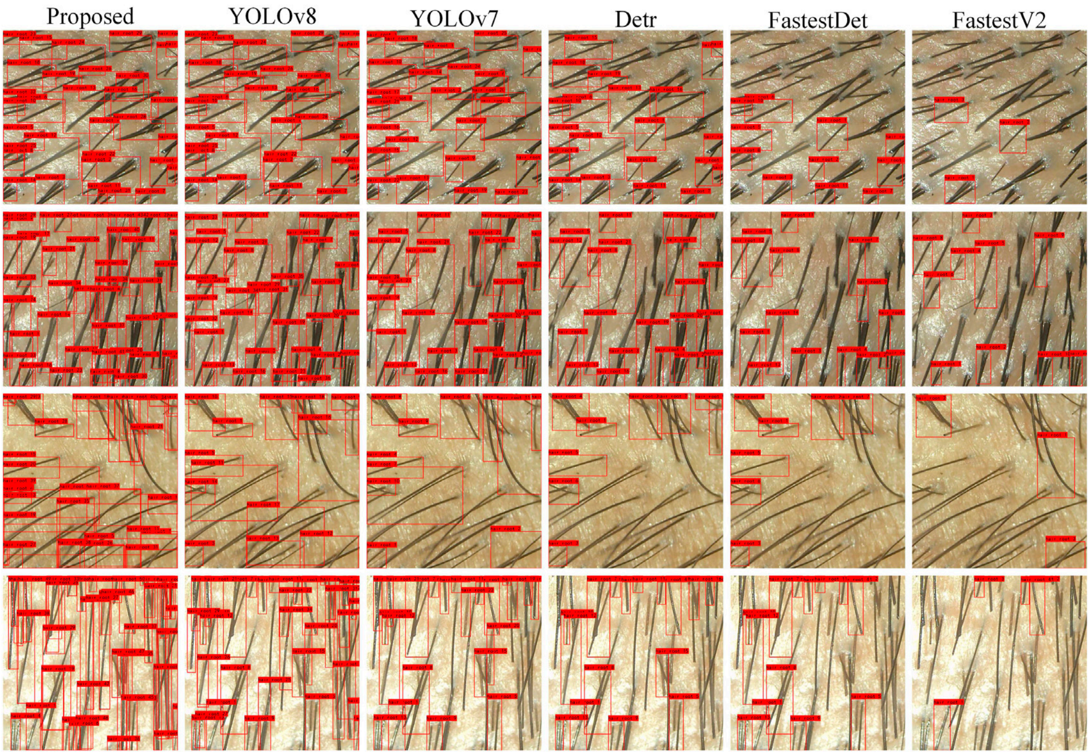


FIGURE 6
Visual comparison of hair cluster detection results.

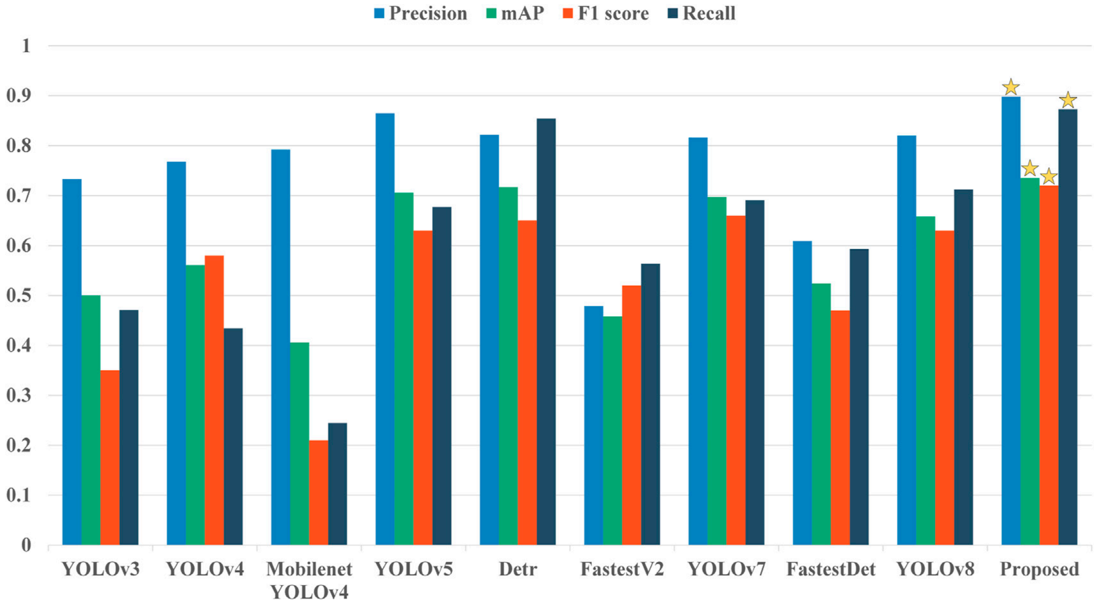


FIGURE 7
Performance comparison of different detection methods on the four indicators of Precision, Recall, mAP (mean average precision), and F1 score. The method that performs best in each case is marked with an asterisk.

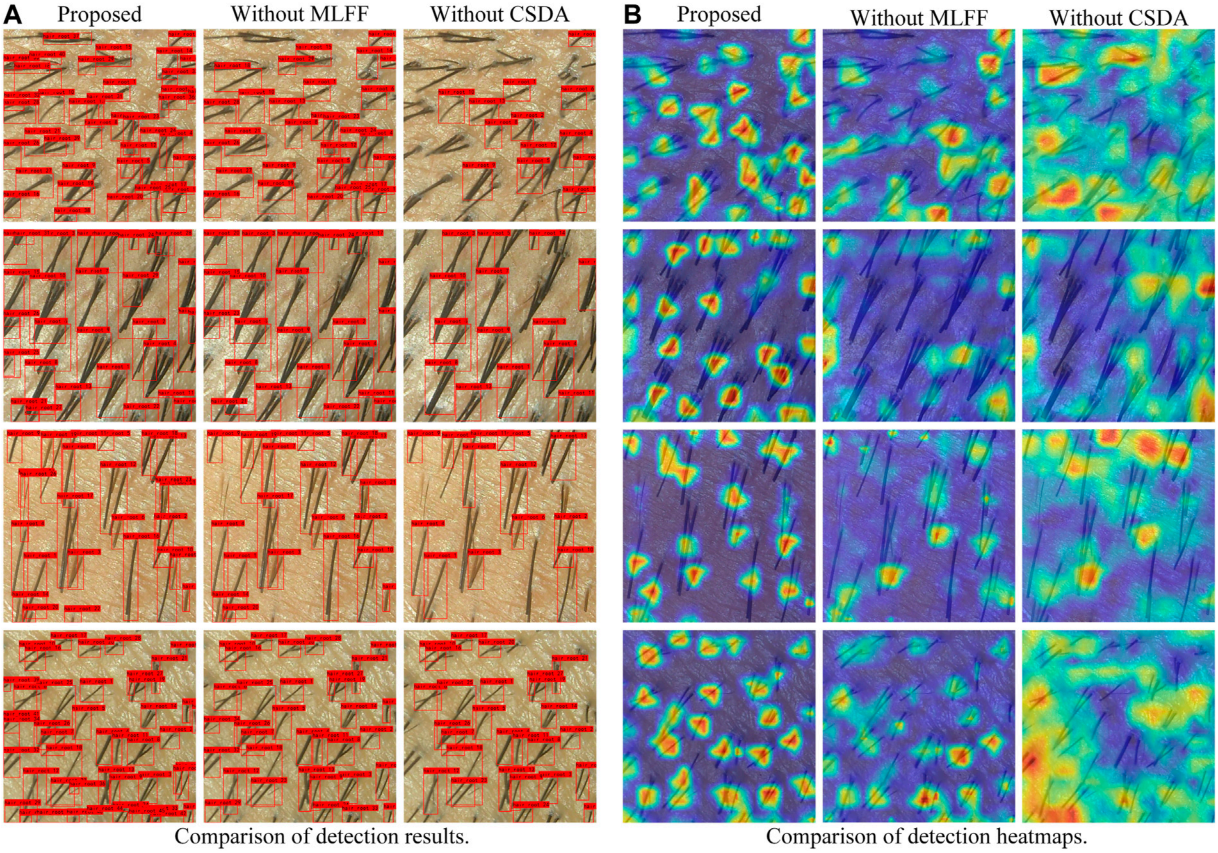


FIGURE 8
Visual comparison of ablation experiment results. **(A)**: Comparison of detection results; **(B)**: Comparison of detection heatmaps.

4.4 Ablation experiment

This study utilizes the developed model as the network for sparse hair target detection (Ours) in hair cluster detection. Experiments were conducted by removing the designed modules from this model. Specifically, the MLFF module was removed from the feature extraction network to assess the extraction of image features, and the CSDA module was removed from the feature enhancement and processing network to examine feature inference and fusion. As shown in the performance metrics results in Table 2, removing the corresponding modules leads to a decrease in the model's detection performance. Additionally, as depicted in Figure 8A, it is apparent that some smaller and overlapping hair clusters are missed when certain modules are removed, while the detection results proposed in this study remain superior.

To further explore the differences between different modules and their reasons, a heatmap analysis was conducted. Figure 8B visualizes the objective performance of different modules. It can be observed that removing the CSDA module generates regions of interest extending beyond the actual target area, focusing on some irrelevant background information. While focusing on certain background regions might not significantly impact normal target detection, it proves detrimental for densely distributed small targets, exacerbating background interference and the difficulty of instance recognition. Without the MLFF module, the situation of missed detections is more severe, indicating that the inclusion of the MLFF module in the network brings more information about the target. In conclusion, the proposed modules in this study contribute to improving the model's detection performance to a certain extent, significantly enhancing the overall performance of the target detection network.

5 Conclusion

In this study, we have proposed and implemented an efficient and accurate detection model specifically designed for sparse hair clusters. This model is based on an improved neural network for object detection. The construction of this model introduces three innovative aspects: firstly, we designed a new neural network structure based on existing advanced object detection networks to optimize the detection of sparse hair. Secondly, a novel multi-level feature fusion structure was devised to better extract and fuse features at different levels. Lastly, a new attention mechanism, the Channel-Spatial Bi-Attention Module, was introduced to simultaneously consider information in both channel and spatial dimensions, further enhancing the model's expressive power and the accuracy of sparse hair detection.

The model primarily consists of three parts: a feature extraction backbone network, a feature enhancement and processing network, and a detection network. It effectively achieves the detection of hair clusters, predicting the number of hair clusters with promising results in experiments. Despite the application of dermoscopy in hair detection being in an exploratory and developing stage, and related research being incomplete, our study provides a new and effective tool for the precise detection of sparse hair clusters. It opens up new avenues for research and applications in hair detection, contributing to the advancement of dermoscopy in hair detection. This, in turn, assists healthcare professionals in diagnosing conditions and selecting treatment plans, while also providing convenience for daily management and condition monitoring for individuals with hair loss.

If the decisions made by the model are not interpretable, they may not be accepted by individuals. In future research, our project team will explore the interpretability of the hair cluster object detection network, applying these advancements to help healthcare professionals understand the processes in image analysis. Additionally, in order to bring the detection model to edge devices for user convenience, we will explore the development of lightweight hair cluster object detection models in the future.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Medical Ethics Committee of the Second Affiliated Hospital of Army Medical University of Chinese People's Liberation Army. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

YX: Data curation, Software, Supervision, Visualization, Writing–original draft. KY: Resources, Software, Validation, Writing–original draft. YL: Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Software, Writing–review and editing. ZL: Data curation, Formal Analysis, Project administration, Software, Supervision, Writing–review and editing. DF: Conceptualization, Data curation, Resources, Writing–original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Sperling LC, Mezebish DS. Hair diseases. *Med Clin North America* (1998) 82: 1155–69. doi:10.1016/s0025-7125(05)70408-9

2. Franzoi SL, Anderson J, Frommelt S. Individual differences in men's perceptions of and reactions to thinning hair. *J Soc Psychol* (1990) 130:209–18. doi:10.1080/00224545.1990.9924571

3. Shapiro J. Hair loss in women. *New Engl J Med* (2007) 357:1620–30. doi:10.1056/nejmcp072110

4. Ahmed A, Almohanna H, Griggs J, Tosti A. Genetic hair disorders: a review. *Dermatol Ther* (2019) 9:421–48. doi:10.1007/s13555-019-0313-2

5. York K, Meah N, Bhoyrul B, Sinclair R. A review of the treatment of male pattern hair loss. *Expert Opin Pharmacother* (2020) 21:603–12. doi:10.1080/14656566.2020.1721463

6. O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, et al. Deep learning vs traditional computer vision. In Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1; 25-26 April 2019; Las Vegas, Nevada, USA. Springer (2020). 128–44.

7. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E, et al. Deep learning for computer vision: a brief review. *Comput intelligence Neurosci* (2018) 2018:1–13. doi:10.1155/2018/7068349

8. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. *NPJ digital Med* (2021) 4:5. doi:10.1038/s41746-020-00376-2

9. Li W, Raj ANJ, Tjahjadi T, Zhuang Z. Digital hair removal by deep learning for skin lesion segmentation. *Pattern Recognition* (2021) 117:107994. doi:10.1016/j.patcog.2021.107994

10. Attia M, Hossny M, Zhou H, Nahavandi S, Asadi H, Yazdabadi A. Digital hair segmentation using hybrid convolutional and recurrent neural networks architecture. *Comp Methods Programs Biomed* (2019) 177:17–30. doi:10.1016/j.cmpb.2019.05.010

11. Kim M, Gil Y, Kim Y, Kim J. Deep-learning-based scalp image analysis using limited data. *Electronics* (2023) 12:1380. doi:10.3390/electronics12061380

12. Hosny KM, Elshora D, Mohamed ER, Vrochidou E, Papakostas GA. Deep learning and optimization-based methods for skin lesions segmentation: a review. *IEEE Access* (2023) 11:85467–88. doi:10.1109/access.2023.3303961

13. Nam G, Wu C, Kim MH, Sheikh Y. Strand-accurate multi-view hair capture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 16 2019 to June 17 2019; Long Beach, CA, USA (2019). p. 155–64.

14. Cuéllar F, Puig S, Kolm I, Puig-Butille J, Zaballos P, Martí-Laborda R, et al. Dermoscopic features of melanomas associated with mc1r variants in Spanish cdkn2a mutation carriers. *Br J Dermatol* (2009) 160:48–53. doi:10.1111/j.1365-2133.2008.08826.x

15. Tosti A, Torres F. Dermoscopy in the diagnosis of hair and scalp disorders. *Actas dermo-sifiliográficas* (2009) 100:114–9. doi:10.1016/s0001-7310(09)73176-x

16. Pirmez R, Tosti A. Trichoscopy tips. *Dermatol Clin* (2018) 36:413–20. doi:10.1016/j.det.2018.05.008

17. Van Camp YP, Van Rompaey B, Elseviers MM. Nurse-led interventions to enhance adherence to chronic medication: systematic review and meta-analysis of randomised controlled trials. *Eur J Clin Pharmacol* (2013) 69:761–70. doi:10.1007/s00228-012-1419-y

18. Shen X, Yu RX, Shen CB, Li CX, Jing Y, Zheng YJ, et al. Dermoscopy in China: current status and future prospective. *Chin Med J* (2019) 132:2096–104. doi:10.1097/cm9.0000000000000396

19. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. *Inf Fusion* (2023) 91: 376–87. doi:10.1016/j.inffus.2022.10.022

20. He X, Qi G, Zhu Z, Li Y, Cong B, Bai L. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. *Simulation Model Pract Theor* (2023) 126:102769. doi:10.1016/j.simpat.2023.102769

21. Zhou F, Zhao H, Nie Z. Safety helmet detection based on yolov5. In: 2021 IEEE International conference on power electronics, computer applications (ICPECA) (IEEE); January 22-24, 2021; Shenyang, China (2021). p. 6–11.

22. Huang Z, Jiang X, Wu F, Fu Y, Zhang Y, Fu T, et al. An improved method for ship target detection based on yolov4. *Appl Sci* (2023) 13:1302. doi:10.3390/app13031302

23. Qi G, Wang H, Haner M, Weng C, Chen S, Zhu Z. Convolutional neural network based detection and judgement of environmental obstacle in vehicle operation. *CAAI Trans Intelligence Tech* (2019) 4:80–91. doi:10.1049/trit.2018.1045

24. Qi G, Zhang Q, Zeng F, Wang J, Zhu Z. Multi-focus image fusion via morphological similarity-based dictionary construction and sparse representation. *CAAI Trans Intelligence Tech* (2018) 3:83–94. doi:10.1049/trit.2018.0011

25. Li Y, Wang Z, Yin L, Zhu Z, Qi G, Liu Y. X-net: a dual encoding–decoding method in medical image segmentation. *Vis Comp* (2021) 39:2223–33. doi:10.1007/s00371-021-02328-7

26. Lama N, Kasmi R, Hagerty JR, Stanley RJ, Young R, Miinch J, et al. Chimeranet: U-net for hair detection in dermoscopic skin lesion images. *J Digital Imaging* (2023) 36: 526–35. doi:10.1007/s10278-022-00740-6

27. Sacha JP, Caterino TL, Fisher BK, Carr GJ, Youngquist RS, D'Alessandro BM, et al. Development and qualification of a machine learning algorithm for automated hair counting. *Int J Cosmet Sci* (2021) 43:S34–S41. S34–S41. doi:10.1111/ics.12735

28. Yoon HS, Park SW, Yoo JH. Real-time hair segmentation using mobile-unet. *Electronics* (2021) 10:99. doi:10.3390/electronics10020099

29. Wu W, Liu H, Li L, Long Y, Wang X, Wang Z, et al. Application of local fully convolutional neural network combined with yolo v5 algorithm in small target detection of remote sensing image. *PloS one* (2021) 16:e0259283. doi:10.1371/journal.pone.0259283

30. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 18 2018 to June 23 2018; Salt Lake City, UT, USA (2018). 7132–41.

31. Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L, et al. Dynamic head: unifying object detection heads with attentions. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; June 20 2021 to June 25 2021; Nashville, TN, USA (2021). 7373–82.

32. Redmon J, Farhadi A. Yolov3: an incremental improvement[J] (2018). arXiv preprint arXiv:1804.02767 Available at: https://arxiv.org/pdf/1804.02767.pdf.

33. Bochkovskiy A, Wang CY, Liao HYM. Yolov4: optimal speed and accuracy of object detection[J] (2020). arXiv preprint arXiv:2004.10934 Available at: https://arxiv.org/abs/2004.10934.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 19 February 2024
DOI 10.3389/fphy.2024.1371320

Check for updates

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Yunze Wang,
Shijiazhuang Tiedao University, China
Zhe Li,
Hunan University, China
Ye Li,
Central South University, China

*CORRESPONDENCE
Yanqiu Bi,
✉ biyanqiu@cqjtu.edu.cn

RECEIVED 16 January 2024
ACCEPTED 07 February 2024
PUBLISHED 19 February 2024

CITATION
Luo Z, Bi Y, Yang X, Li Y, Yu S, Wu M and Ye Q (2024), Enhanced YOLOv5s + DeepSORT method for highway vehicle speed detection and multi-sensor verification.
Front. Phys. 12:1371320.
doi: 10.3389/fphy.2024.1371320

Enhanced YOLOv5s + DeepSORT method for highway vehicle speed detection and multi-sensor verification

Zhongbin Luo[1,2], Yanqiu Bi[3,4]*, Xun Yang[1,2], Yong Li[5,6], Shanchuan Yu[1,2], Mengjun Wu[1,2] and Qing Ye[1,2]

[1]China Merchants Chongqing Communications Research and Design Institute Co., Ltd., Chongqing, China, [2]Research and Development Center of Transport Industry of Self-Driving Technology, Chongqing, China, [3]National and Local Joint Engineering Research Center of Transportation Civil Engineering Materials, Chongqing Jiaotong University, Chongqing, China, [4]School of Civil Engineering, Chongqing Jiaotong University, Chongqing, Shandong, China, [5]College of Computer Science, Chongqing University, Chongqing, China, [6]Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing, China

Addressing the need for vehicle speed measurement in traffic surveillance, this study introduces an enhanced scheme combining YOLOv5s detection with Deep SORT tracking. Tailored to the characteristics of highway traffic and vehicle features, the dataset data augmentation process was initially optimized. To improve the detector's recognition capabilities, the Swin Transformer Block module was incorporated, enhancing the model's ability to capture local regions of interest. *CIoU* loss was employed as the loss function for the vehicle detection network, accelerating model convergence and achieving higher regression accuracy. The Mish activation function was utilized to reduce computational overhead and enhance convergence speed. The structure of the Deep SORT appearance feature extraction network was modified, and it was retrained on a vehicle re-identification dataset to mitigate identity switches due to obstructions. Subsequently, using known references in the image such as lane markers and contour labels, the transformation from image pixel coordinates to actual coordinates was accomplished. Finally, vehicle speed was measured by computing the average of instantaneous speeds across multiple frames. Through radar and video Multi-Sensor Verification, the experimental results show that the mean Average Precision (mAP) for target detection consistently exceeds 90%. The effective measurement distance for speed measurement is around 140 m, with the absolute speed error generally within 1–8 km/h, meeting the accuracy requirements for speed measurement. The proposed model is reliable and fully applicable to highway scenarios.

KEYWORDS

YOLOv5S, Deep SORT, swin transformer, vehicle speed, traffic monitoring

1 Introduction

Intelligent Transportation Systems (ITS) have been widely applied to practical traffic scenarios such as highways, urban roads, tunnels, and bridges. This integration owes much to the convergence of various technologies, including pattern recognition, video image processing, and network communication [1, 2]. Vehicle speed is a crucial parameter that directly reflects the state of traffic [3, 4]. Meanwhile, in highly complex traffic monitoring scenarios and under special weather conditions, intelligent transportation monitoring systems face numerous significant challenges. In addressing the issue of vehicle speeding, the measurement of vehicle speed can provide vital data for traffic management authorities. Accurate measurement of vehicle target speed is one of the challenges faced by traffic monitoring systems.

Traditional vehicle speed detection primarily utilizes inductive loop detection, laser detection, and radar detection. These methods are well-developed and commonly used in traffic systems. However, traditional detection methods have the following disadvantages: (1) the required equipment is expensive; (2) the equipment is installed under the road surface, leading to high subsequent maintenance costs and maintenance not only affects traffic but also damages road structure. Video-based vehicle speed detection leverages numerous traffic video monitoring devices, significantly overcoming the high costs and difficult maintenance issues associated with traditional speed detection methods. The vehicle speed detection system can be categorized into two types: one type focuses on accurate speed monitoring systems (such as speed camera applications) [5, 6], and the other type, though less precise, can be used to estimate traffic speed (such as traffic camera application scenarios) [7, 8]. This classification system takes into account the intrinsic parameters of the camera (such as sensor size and resolution, focal length), as well as extrinsic parameters (such as the camera's position relative to the road surface, drone-based cameras, etc.), and the number of cameras (monocular, stereo, or multiple cameras).

Through these parameters, the actual scene on the image plane can represent one or multiple lanes, as well as the relative position of vehicles to the camera, ultimately yielding one of the most critical variables: the ratio of pixels to road segment length, i.e., the road length each pixel represents. Due to the perspective projection model, this ratio is directly proportional to the square of the camera's distance, implying that measurements over long distances have poor accuracy. Accurate estimation of the camera's intrinsic and extrinsic parameters is required to provide measurements in the actual coordinate system. The most common approach is soft calibration, which involves calibrating intrinsic parameters in a verification laboratory or using sensor and lens characteristics, and estimating the rigid transformation between the camera and the road surface using manual [9, 10] or automatic [11] methods.

Hard calibration involves estimating both the intrinsic and extrinsic parameters of the camera, which can be done either manually [12] or automatically [13–15]. In certain limited scenarios, some details of camera calibration may be overlooked, such as the exact position of the camera, anchoring systems, gantries. Since cameras are mostly static (except for drone cameras), vehicle detection is most often addressed by modeling the background [16–18]. Other methods are feature-based, such as detecting vehicle license plates [19, 20] or other characteristics [21–23].

Recently, learning-based approaches have become increasingly popular for recognizing vehicles in images [24, 25]. The ability to track vehicles with smooth and stable trajectories is a key issue in handling vehicle speed detection. Vehicle tracking can be divided into three different categories: The first category is feature-based [26–28], where tracking originates from a set of features of the vehicle (such as optical flow). The second category focuses on tracking the centroid of a vehicle's blob or bounding box [29, 30]. The third category concentrates on tracking the entire vehicle [31, 32] or its specific parts (such as the license plate [33, 34]).

The prerequisite for speed measurement is the effective assessment of distance. In monocular vision systems, the estimation of vehicle distance typically relies on specific constraints and methods. These include: (1) Flat road assumption and homography-based methods, which assume that the road is flat and apply a mathematical transformation known as homography [35, 36], helping in mapping the view of a scene from one perspective to another, which is crucial for estimating distances in 2D images; (2) Detection of lines and specific areas [37, 38]. By detecting lines and specific areas, designed detection lines and areas can be overlaid on the real-world view, providing a reference scale for measuring distances; (3) Use of prior knowledge about object dimensions, utilizing the known dimensions of certain objects to estimate distances. For instance, knowing the standard sizes of license plates ([39, 40]) or the average dimensions of vehicles [41] can assist in calibrating distance measurements. However, these monocular methods have limitations, which are addressed in stereo vision systems. In stereo vision systems [42], two cameras are used to capture the same scene from slightly different angles, similar to human binocular vision. This setup allows for more accurate depth perception and distance estimation, as it mimics the way.

Currently, speed detection is primarily divided into macroscopic traffic flow speed and individual vehicle speed. Macroscopic traffic flow speed detection is based on a specific road section, using the length of the section and travel time to estimate the average speed of the segment [43, 44]. Individual vehicle speed detection focuses on the micro-level speed of the vehicle itself, presenting greater technical challenges. This process requires prior knowledge of the camera's frame rate or accurate timestamps for each image to calculate the time between measurements. Utilizing consecutive or non-consecutive [45] images to estimate speed is a key factor impacting accuracy. In summary, whether in traffic flow speed or individual vehicle speed detection, factors such as the method of image capture (continuous or non-continuous), frame rate, timestamp accuracy, and the integration of various measurement data need to be carefully considered. The selection method and precision of these factors directly affect the accuracy of speed estimation.

In summary, vision-based vehicle speed detection involves the entire process of camera calibration, distance estimation, and speed estimation. However, the calibration process for monocular vision cameras is complex, the accuracy of distance estimation is relatively poor, and the precision of individual vehicle speed estimation needs improvement. Currently, there are few instances of rapidly detecting and stably tracking vehicle instantaneous speeds solely through video recognition technology, which limits the broader application of video recognition technologies in the field of traffic safety. Therefore, this study introduces an enhanced scheme that combines YOLOv5s detection with Deep SORT tracking, targeting the need for vehicle speed measurement in traffic monitoring. The dataset data expansion process is preliminarily optimized based on the characteristics of highway traffic and vehicle features. The Swin Transformer Block module is introduced to improve the detector's

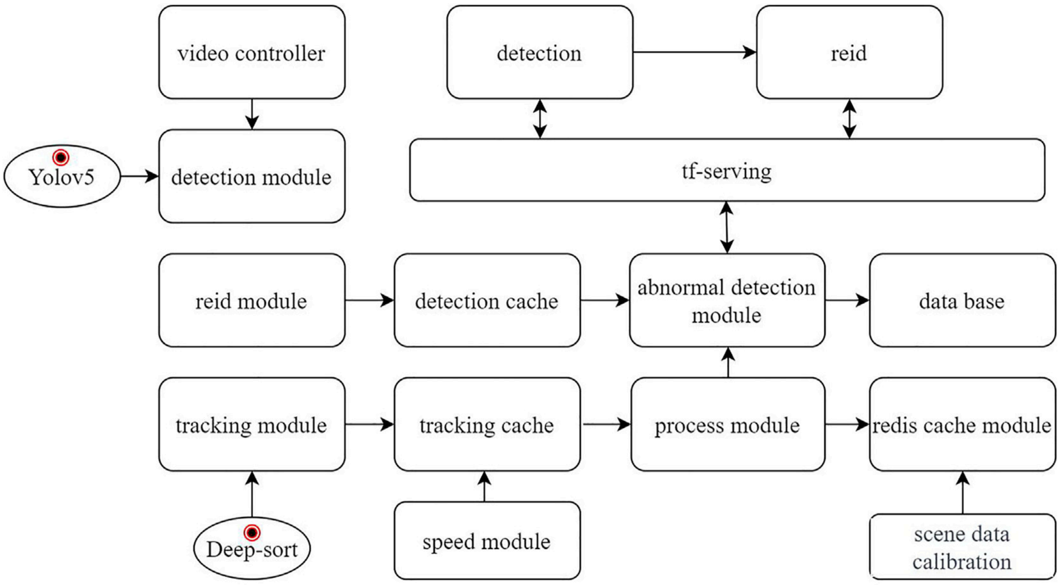


FIGURE 1
Algorithm framework diagram.

recognition capabilities and enhance the model's ability to capture areas of interest. The *CIoU* loss is employed as the loss function for the vehicle detection network to accelerate model convergence and achieve higher regression precision. The Mish activation function is used to reduce computational costs and improve convergence speed. Modifications are made to the structure of the Deep SORT appearance feature extraction network, and it is retrained on the vehicle re-identification dataset to mitigate identity switches caused by obstacles. Subsequently, known references in the image, such as lane markings and contour labels, are used to complete the conversion from image pixel coordinates to actual coordinates through maximum likelihood estimation, maximum posterior estimation, and non-linear least squares methods. Finally, vehicle speed is measured by calculating the average of instantaneous speeds over multiple frames. The algorithm can detect and track vehicle targets without prior camera parameters and calibration, extract known reference information such as lane lines and contour labels, and automatically convert pixel coordinates to actual coordinates in traffic monitoring scenes, as well as automatically measure vehicle speeds, the algorithm framework as shown in Figure 1. Accurate estimation of vehicle speed can support the detection of traffic accidents and incidents, offering scientific technical means for active safety management in intelligent transportation systems.

2 Improved YOLOv5s + DeepSORT algorithm for highway vehicle detection and tracking

2.1 Construction of vehicle target dataset

2.1.1 Characteristics of highway traffic scenarios

There are typically four categories of common highway traffic scenarios, as shown in Figure 2.

(a) Scene variations, as the setup of traffic monitoring varies, so do the monitoring angles and heights. For instance, the monitoring angle and scene characteristics inside a tunnel differ greatly from those on a highway, leading to significantly reduced detection accuracy and numerous false detections of vehicle targets, as shown in Figure 2A.

(b) The same scene at different times also exhibits significant differences. With changes in time, the brightness and visibility of scene images vary. The characteristics of vehicle targets at night are particularly difficult to capture due to the substantial interference from vehicle lights at night, making it hard to accurately obtain the body contours of target vehicles. If the dataset does not include such special night scene data, the detection results are not ideal Figure 2B.

(c) Vehicle targets at different positions in the image will have obvious deformation. The same vehicle target will undergo significant size deformation from distant to closer positions in the image, affecting the detection accuracy of small targets. The red boxes in Figure 2C indicate significant deformations of the same vehicle target at different locations.

(d) On actual roads, there is a widespread occurrence of vehicle occlusion, which can lead to multiple targets being detected as one, resulting in missed and false detections. The red boxes in Figure 2D represent situations where vehicles are obstructing each other.

The existence of these four types of issues makes large public datasets such as COCO and VOC unsuitable for the perspectives captured by highway cameras, leading to a large number of false positives and missed detections of target vehicles.

2.1.2 Data preparation

Given the relatively uniform types of motor vehicles in highway scenarios, vehicles are generally classified into three

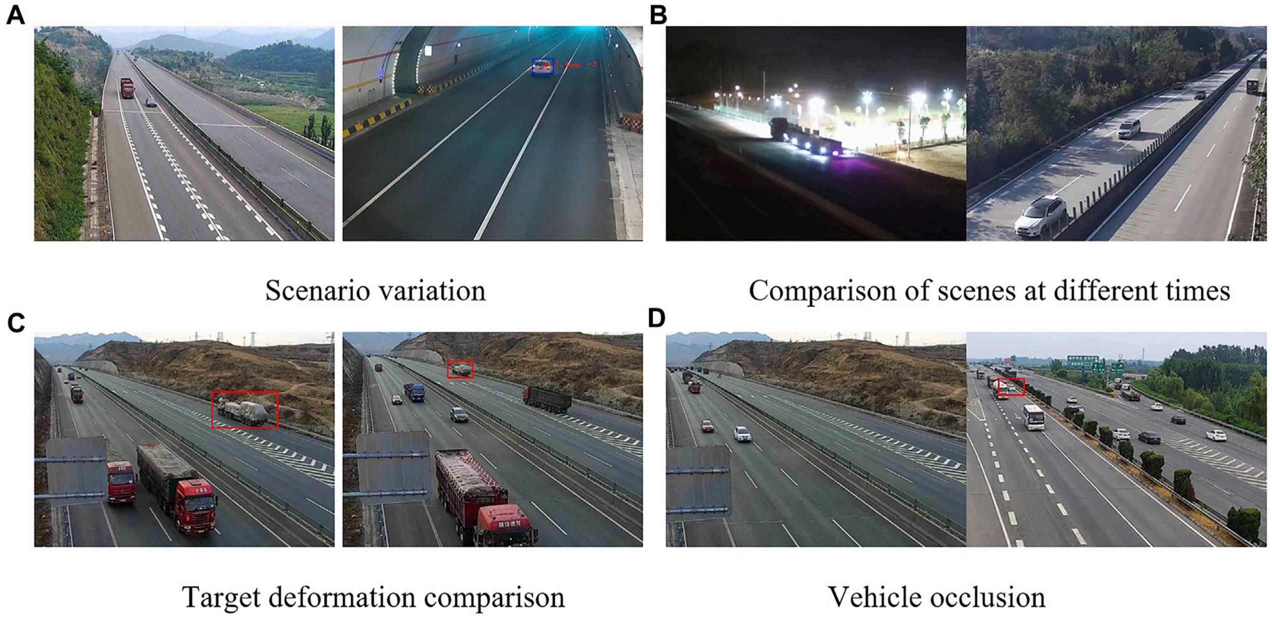

A
Scenario variation
B
Comparison of scenes at different times
C
Target deformation comparison
D
Vehicle occlusion

FIGURE 2
Common issues in target vehicle detection.

categories: Car, Bus, and Truck. Car mainly refer to passenger vehicles with seating for fewer than seven people; Bus mainly include commercial buses, public transport buses, etc.; Truck primarily refer to small, medium, and large trucks, trailers, and various types of special-purpose vehicles as shown in Table 1. By collecting datasets from different scenes on highways and manually labeling them using the labelImg tool, a dataset in YOLO format was ultimately created.

The specific process includes: (1) Data Collection: Collect representative image data covering various scenes and angles of target categories. (2) Data Division: Divide the dataset into training, validation, and test sets, typically in a certain ratio, to ensure the independence and generalizability of the data. (3) Bounding Box Annotation: Annotate each target object with a bounding box, usually represented by a rectangle, including the coordinates of the top-left and bottom-right corners. Category Labeling: Assign corresponding category labels to each target object, identifying the category to which the object belongs. During dataset annotation, rectangular bounding boxes encompassing the entire vehicle are marked, with each side fitting closely to the vehicle. Annotation is not performed when the occlusion exceeds 50%, the vehicle type is indistinguishable, or the size is below 10*10 pixels. Furthermore, in cases where vehicles are truncated, the truncation is not considered to affect the overall annotation. Trucks used for transportation are uniformly annotated, without separately marking the vehicles on them.

2.1.3 Data augmentation

To enhance the accuracy and generalization capability of model training, data augmentation techniques are employed, tailored to the characteristics of highway traffic environments and vehicle features. These techniques include Mosaic, Random_perspective, Mixup, HSV, Flipud, Fliplr, as shown in Figure 3.

2.2 Optimization of object detection network

In response to the identified issues with YOLOv5 in highway vehicle detection, the following optimizations were made to enhance the accuracy of vehicle detection: (1) Incorporating the Swin Transformer Block module to improve the model's ability to capture information from local areas of interest; (2) Utilizing *CIoU* loss as the loss function for the vehicle detection network to accelerate model convergence and achieve higher regression accuracy; (3) Adopting the Mish activation function to reduce computational overhead and increase convergence speed.

2.2.1 Introduction of swin transformer block

To address the shortcomings of traditional YOLOv5 in traffic object detection, the Swin Transformer Block module is introduced for optimization.

The Swin Transformer network [46], proposed in 2021, is a Transformer network enhanced with a local self-attention mechanism. It has stronger dynamic computation capabilities compared to convolutional neural networks, with enhanced modeling capacity, and can adaptively compute both local and global pixel relationships, making it highly valuable for widespread use.

The core modules of the Transformer Block overall architecture are the Window-based Multi-Head Self-Attention layer (W-MSA) and the Shifted Window-based Multi-Head Self-Attention layer (SW-MSA). By restricting attention computation within a window, the network not only introduces the locality of convolution operations but also saves computational resources, resulting in good performance.

This article proposes the integration of the Swin Transformer Block structure into the backbone feature extraction network and

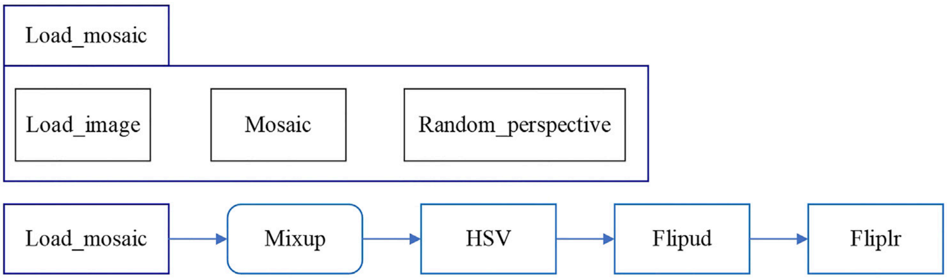


FIGURE 3
Data augmentation Flowchart.

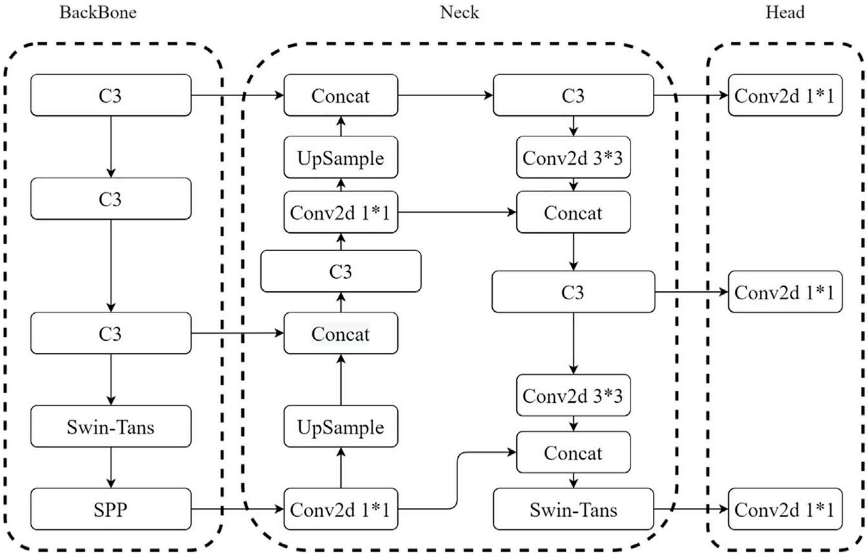


FIGURE 4
SwinTransYOLOv5 network structure diagram.

neck feature fusion, utilizing the efficient self-attention mechanism module to fully explore the potential of feature representation. The improved YOLOv5 network incorporating the Swin Transformer Block module is shown in Figure 4, named SwinTransYOLOv5 network.

2.2.2 Improvement of loss function

YOLOv5s employs $GIoU$ loss as the bounding box regression loss function to evaluate the distance between the predicted bounding box (PB) and the ground truth bounding box (GT), as shown in Eq. 1.

$$\begin{cases} GIoU = IoU - \dfrac{A^c - U}{A^c} \\ L_{GIoU} = 1 - GIoU \end{cases} \tag{1}$$

In the formula, IoU represents the intersection over union of PB and GT, A^c is the area of the smallest rectangular box containing both PB and GT, U is the union of PB and GT, and L_{GIoU} is the $GIoU$ loss. The advantage of $GIoU$ loss is its scale invariance, meaning the similarity between PB and GT is independent of their spatial scale. The problem with $GIoU$ Loss is that when either PB or GT completely encompasses the other, $GIoU$ Loss degenerates entirely into IoU loss. Because it heavily relies on the IoU term, this results in slow convergence during actual training and lower accuracy of the predicted bounding boxes. To address these issues, $CIoU$ loss also considers the overlapping area of PB and GT, the distance between their centroids, and their aspect ratios, as shown in Eq. 2.

$$\begin{cases} CIoU = IoU - \dfrac{\rho^2(b, b^{gt})}{c^2} - av \\ L_{CIoU} = 1 - CIoU \end{cases} \tag{2}$$

In the formula, b and b^{gt} represent the centroids of PB and GT, $\rho^2(.)$ denotes the Euclidean distance, c is the length of the shortest diagonal of the smallest enclosing box of PB and GT, a represents a positive balance parameter, and v indicates the consistency of the aspect ratio of PB and GT. The definitions of a and v are as follows in Eq. 3.

TABLE 1 Dataset categorization.

Type	Example
Car	
Bus	
Truck	

$$\begin{cases} v = \dfrac{4}{\pi^2}\left(arctan\dfrac{\omega^{gt}}{h^{gt}} - arctan\dfrac{\omega}{h}\right) \\ a = \dfrac{v}{(1 - IoU) + v} \end{cases} \tag{3}$$

In the formula, ω^{gt}, h^{gt} and ω, h respectively represent the width and height of GT and PB.

Compared to the *GIoU* loss used in YOLOv5s, *CIoU* loss incorporates penalty terms for the distance between the centers of PB and GT, as well as their aspect ratios in the loss function. This ensures faster convergence of the predicted bounding boxes during training and yields higher regression localization accuracy. In This article, *CIoU* loss is adopted as the loss function for the vehicle detection network.

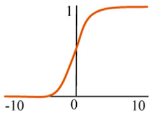
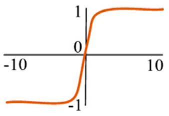
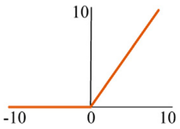
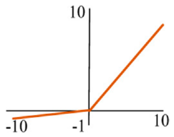
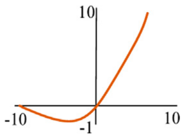
2.2.3 Activation function

Changing the activation function can significantly enhance recognition performance. Activation functions are categorized into saturated and non-saturated types. The primary advantages of using non-saturated activation functions are twofold [47]: firstly, they effectively address the vanishing gradient problem, which becomes more severe with saturated activation functions; secondly, they can accelerate the convergence speed. After comparing the pros, cons, and characteristics of various activation functions without significantly increasing computational load, as shown in Table 2, the Leaky ReLU activation function in YOLOv5 was replaced with the Mish activation function.

2.3 Optimization of deep SORT for vehicle tracking

The multi-object online tracking algorithm SORT [48] (Simple Online and Realtime Tracking) utilizes Kalman filtering and Hungarian matching, using the *IoU* between tracking and detection results as the cost matrix, to implement a simple, efficient, and practical tracking paradigm. However, the SORT algorithm's limitation lies in its association metric being effective only when the uncertainty in state estimation is low, leading to

TABLE 2 Comparison of common activation functions.

	Sigmoid	tanh	ReLU	Leaky ReLU	Mish
Function graphs					
Function Formula	$\delta(x) = \frac{1}{1+e^{-x}}$	$tanh(x)$	$max(0, x)$	$max(0.1x, x)$	$x^{*} tanh(softflus(x))$
Advantages	Can restrict the output to be between (0, 1), facilitating the completion of classification tasks	①Can restrict the output to be between (−1, 1), facilitating the completion of classification tasks ②Zero-Centered	Linear: Saves computational resources and shortens convergence time	①Linear ②Gradient non-saturation, no neuron death	①Linear ②Gradient non-saturation, no neuron death ③The network's convergence is the best among the five activation functions
Disadvantages	①The output is not zero-centered, leading to a zigzag pattern in gradient descent ②Gradient saturation, Gradient vanishing ③Non-linear, involves exponential operations, consuming more resources during computation	②Gradient saturation, Gradient vanishing ③Non-linear	Neuron Death: The left side of the ReLU function is completely flat. When the neuron's z-value is negative, the output α is 0, and the gradient is also 0, making it impossible to alter the weight value w through the gradient, leaving w unchanged	The network's convergence is not advantageous compared to the latest networks	Relatively higher computational cost

numerous identity switches and tracking failures when the target is occluded. To address this issue, Deep SORT [49] combines both motion and appearance information of the target as the association metric, improving tracking failures caused by the target's disappearance and reappearance.

2.3.1 Tracking processing and state estimation

Deep SORT uses an 8-dimensional state space (u, v, y, h, x, y, y, h) to describe the target's state and motion information in the image coordinate system. u and v represent the center coordinates of the target detection box, y and h respectively represent the aspect ratio and height of the detection box, and (x, y, y, h) represent the relative velocity of the previous four parameters in the image coordinates. The algorithm employs a standard Kalman filter with a constant velocity model and a linear observation model, using the detection box parameters (u, v, y, h) as direct observations of the object state. By combining motion and appearance information, the Hungarian algorithm is used to match predicted and tracked boxes, and cascaded matching is integrated to enhance accuracy.

(1) Mahalanobis Distance

The Mahalanobis distance is used to evaluate the predicted Kalman state and the new state, as shown in Eq. 4.

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i) \tag{4}$$

$d^{(1)}(i, j)$ represents the motion matching degree between the j detection and the i trajectory, where S_i is the covariance matrix of the observation space at the current moment predicted by the Kalman filter for the trajectory, y_i is the predicted observation of the trajectory at the current moment, and d_j is the state of the j detection.

Considering the continuity of motion, detections are filtered using this Mahalanobis distance, with the 0.95 quantile of the chi-square distribution as the threshold value, defining a threshold function, as shown in Eq. 5.

$$b_{i,j}^{(1)} = 1\left[d^{(1)}(i,j) \le t^{(1)}\right] \tag{5}$$

(2) Appearance features

While Mahalanobis distance is a good measure of association when the target's motion uncertainty is low, it becomes ineffective in practical situations like camera movement, leading to a large number of mismatches. Therefore, we integrate a second metric. For each BBox detection, we compute an appearance feature descriptor. We create a gallery to store the descriptors of the latest 100 trajectories and then use the minimum cosine distance between the i and j trajectories as the second measure, as shown in Eq. 6.

$$d^{(2)}(i,j) = min\left\{1 - r_j{}^T r_k^{(i)} \mid r_k^{(i)} \in \mathcal{R}_i\right\} \tag{6}$$

Can be represented using a threshold function, as shown in Eq. 7.

TABLE 3 Adjusted reconstruction network.

Network layer	Convolutional kernel parameters	Output size
Conv 1	3 × 3/1	32 × 128×128
Conv 2	3 × 3/1	32 × 128×128
Max Pool 3	3 × 3/2	32 × 64×64
Residual 4	3 × 3/1	32 × 64×64
Residual 5	3 × 3/1	32 × 64×64
Residual 6	3 × 3/2	64 × 32×32
Residual 7	3 × 3/1	64 × 32×32
Residual 8	3 × 3/2	128 × 16×16
Residual 9	3 × 3/1	128 × 16×16
Dense 10	-	128
Batch and ℓ_2 Norm	-	128

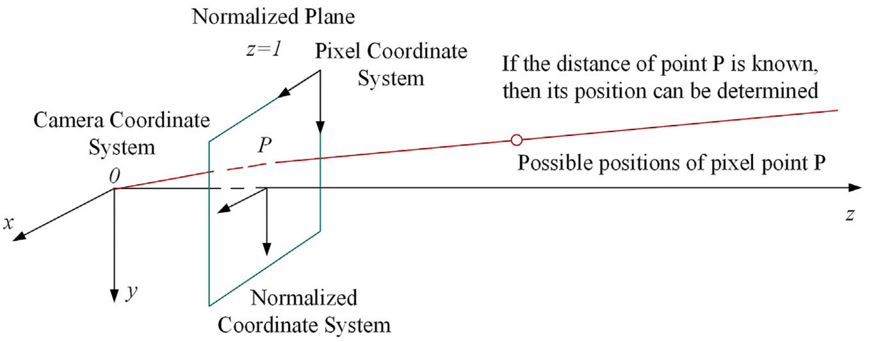

Normalized Plane
z=1
Pixel Coordinate System
If the distance of point P is known, then its position can be determined
Camera Coordinate System
P
0
Possible positions of pixel point P
x
z
y
Normalized Coordinate System

FIGURE 5
Pixel coordinate conversion diagram.

$$b_{i,j}^{(2)} = 1\left[d^{(2)}(i,j) \le t^{(2)}\right] \tag{7}$$

Mahalanobis distance can provide reliable target location information in short-term predictions, and the cosine similarity of appearance features can recover the target ID when the target is occluded and reappears. To make the advantages of both measures complementary, a linear weighting approach is used for their combination, as shown in Eqs 8, 9.

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda)d^{(2)}(i,j) \tag{8}$$

$$b_{i,j} = \prod\nolimits_{m=1}^{2} b_{i,j}^{(m)} \tag{9}$$

In summary, distance measurement is effective for short-term prediction and matching, while appearance information is more effective for matching long-lost trajectories. The choice of hyperparameters depends on the specific dataset. For datasets with significant camera movement, the degree of motion matching is not considered.

(3) Cascaded matching

The strategy of cascaded matching is used to improve matching accuracy, mainly because when a target is occluded for a long time, the uncertainty of Kalman filtering greatly increases, leading to a dispersion of continuous prediction probabilities. Assuming the original covariance matrix is normally distributed, continuous predictions without updates will increase the variance of this normal distribution, so points far from the mean in Euclidean distance may obtain the same Mahalanobis distance value as points closer in the previous distribution. In the final stage, the authors use *IOU* association from the previous SORT algorithm to match $n = 1$ unconfirmed and unmatched trajectories. This can alleviate significant changes caused by abrupt appearance shifts or partial occlusions. However, this approach may also connect some newly generated trajectories to older ones.

2.3.2 Deep appearance features

The original algorithm uses a residual convolutional neural network to extract the appearance features of the target, training the model on a large-scale pedestrian re-identification dataset for pedestrian detection and tracking. Since the original algorithm was only used for the pedestrian category and the input images were scaled to 128 × 64, which does not match the aspect ratio of vehicle targets, this article improves the network model by adjusting the input image size to 128 × 128, as shown in Table 3. The adjusted network is then re-identification trained on the vehicle re-identification dataset VeRi [50].

3 Vehicle speed measurement

3.1 Model assumptions

All locations in road monitoring images can be mapped to the $Z_w = 0$ plane of the world coordinate system through camera calibration, as shown in Figure 5. However, the precise measurement of vehicle speed depends not only on camera calibration but also significantly on the vehicle's trajectory. To better implement vehicle speed measurement, the speed model assumes the following: (1) In highway scenarios, the road is relatively flat without significant undulations, meeting the condition of $Z_w = 0$; (2) In highway monitoring scenarios, the movement of vehicles between each frame is linear, allowing for the measurement of vehicles moving in both straight and non-straight paths using the proposed speed measurement method; (3) In highway video surveillance, the time interval between each frame is the same, facilitating the calculation of vehicle speed after obtaining the exact vehicle position using the interval between frames.

3.2 Model design and implementation

Based on the assumptions and establishment of the aforementioned speed model, the specific process of speed detection is implemented. Firstly, using the YOLO object detection algorithm, the coordinates of the top-left corner of the image detection box are obtained. By determining the length and width of the detection box, the coordinates of the center of the bottom edge of the box can be obtained. This ensures that the measured vehicle speed is closer to the actual speed. For every target vehicle in each frame of the video stream, a set of vector relations can be obtained, as shown in Eq. 10.

$$\boldsymbol{d_i} = \boldsymbol{u_i}(\boldsymbol{t}) - \boldsymbol{u_i}(\boldsymbol{t} - \Delta \boldsymbol{t}) \tag{10}$$

Here, $u_i(t)$ represents the center coordinates of the bottom edge of the vehicle target detection box in the current video frame; $u_i(t - \Delta t)$ represents the center coordinates of the bottom edge of the vehicle target detection box in the previous frame; Δt is the time interval between the two frames; $i = (1, 2, \ldots, n)$ represents the tracked trajectory points.

d_i represents the pixel distance between adjacent frames, and calculating the speed requires mapping the pixel coordinates to world coordinates. The current common method involves camera calibration, but camera calibration requires knowledge of the camera's focal length, height, internal parameters, etc., and the calibration process can be cumbersome.

In This article, state estimation is performed using the popular methods of maximum likelihood estimation, maximum *a posteriori* estimation, and non-linear least squares, selecting the best estimation parameters based on the loss in state estimation.

(1) Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is an important and widely used method for estimating quantities. MLE explicitly uses a probability model with the goal of finding a system occurrence tree that can produce observed data with a high probability. MLE is a representative of a class of system occurrence tree reconstruction methods based entirely on statistics. Given a set of data, if we know it is randomly taken from a certain distribution, but we don't know the specific parameters of this distribution, that is, "the model is determined, but the parameters are unknown." For example, we know the distribution is a normal distribution, but we don't know the mean and variance; or it's a binomial distribution, but we don't know the mean. MLE can be used to estimate the parameters of the model. The objective of MLE is to find a set of parameters that maximize the probability of the model producing the observed data, as shown in Eq. 11.

$$\underset{\boldsymbol{\mu}}{\mathbf{argmax}}\ \boldsymbol{p}(\boldsymbol{X}; \boldsymbol{\mu}) \tag{11}$$

Here, $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ represents the observed sequence data, and $p(\boldsymbol{X}; \mu)$ is the likelihood function, which denotes the probability of the observed data occurring under the parameter μ. Assuming each observation is independent, as shown in Eq. 12.

$$\boldsymbol{p}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n; \boldsymbol{\mu}) = \prod\nolimits_{i=1}^{n} \boldsymbol{p}(\boldsymbol{x}_i; \boldsymbol{\mu}) \tag{12}$$

To facilitate differentiation, the log is generally taken of the target. Therefore, optimizing the likelihood function is equivalent to optimizing the log-likelihood function, as shown in Eqs 13, 14.

$$\underset{\boldsymbol{\mu}}{\mathbf{arg\,max}}\ \boldsymbol{p}(\boldsymbol{X}; \boldsymbol{\mu}) = \underset{\boldsymbol{\mu}}{\mathbf{arg\,max}}\ \mathbf{log}\,\boldsymbol{p}(\boldsymbol{X}; \boldsymbol{\mu}) \tag{13}$$

$$\boldsymbol{x}^{*}_{\boldsymbol{MLE}} = \mathbf{arg}\,\boldsymbol{maxP}(\boldsymbol{u} \mid \boldsymbol{X}) \tag{14}$$

(2) Maximum A Posteriori Estimation

In Bayesian statistics, Maximum A Posteriori (MAP) Estimation refers to the mode of the posterior probability distribution. MAP estimation is used to estimate the values of quantities that cannot be directly observed in experimental data. It is closely related to the classical method of Maximum Likelihood Estimation (MLE), but it uses an augmented optimization objective that further considers the prior probability distribution of the quantity being estimated. Therefore, MAP estimation can be seen as a regularized form of MLE, as shown in Eqs 15, 16.

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathbf{MAP}} &= \mathbf{arg}\ \underset{\boldsymbol{\theta}}{\boldsymbol{max}}\ \boldsymbol{p}(\boldsymbol{\theta} \mid \boldsymbol{x}) \\ &= \mathbf{arg}\ \underset{\boldsymbol{\theta}}{\boldsymbol{max}}\ \frac{\boldsymbol{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) \times \boldsymbol{p}(\boldsymbol{\theta})}{\boldsymbol{P}(\boldsymbol{x})} \\ &= \mathbf{arg}\ \underset{\boldsymbol{\theta}}{\boldsymbol{max}}\ \boldsymbol{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) \times \boldsymbol{p}(\boldsymbol{\theta}) \end{aligned} \tag{15}$$

$$\boldsymbol{x}^{*}_{\boldsymbol{MAP}} = \mathbf{arg}\,\boldsymbol{max}\,\boldsymbol{P}(\boldsymbol{x} \mid \boldsymbol{z}) = \mathbf{arg}\,\boldsymbol{max}\,\boldsymbol{P}(\boldsymbol{z} \mid \boldsymbol{x})\boldsymbol{P}(\boldsymbol{x}) \tag{16}$$

Here, θ is the parameter to be estimated, and $p(\theta \mid x)$ represents the probability of occurrence of x when the estimated parameter is θ.

(3) Non-Linear Least Squares

The Least Squares Method (also known as the Method of Least Squares) is a mathematical optimization technique. It finds the best function match for data by minimizing the sum of the squares of the errors. The Least Squares Method can be used to easily obtain unknown data, ensuring that the sum of the squares of the errors between these obtained data and the actual data is minimized. The Least Squares Method can also be used for curve fitting, and other optimization problems can be expressed using this method by minimizing energy or maximizing entropy. Using the Least Squares Method to estimate the mapping relationship, the mapping parameters are obtained, as shown in Eqs 17, 18.

$$\min_{x} \sum \| y_i - f(x_i) \|^2_{\sum_i^{-1}} \tag{17}$$

Where $f(x_i)$ is a nonlinear function, and $\sum_i^{-1}$ is the covariance matrix.

$$\psi(x) = \sum \| y_i - f(x_i) \|^2_{\sum_i^{-1}} \tag{18}$$

Then, the Gauss-Newton method is used to solve for ψ(x), as shown in Eq. 19:

$$\begin{aligned}\psi(x) &= \sum \| y_i - f(x_i) \|^2_{\sum_i^{-1}} = \sum_{i=1}^{m} \| e_i(x) \|^2 = e_i^T(x) e_i(x) \\ &= \sum_{i=1}^{m} \varphi_i(x)\end{aligned} \tag{19}$$

For the sum of errors, we investigate the *i* term, also performing a second-order Taylor expansion, followed by differentiation. We first calculate its first-order derivative (gradient) and second-order derivative.

First-order derivative, as shown in Eqs 20, 21.

$$\frac{\partial \varphi_i(x)}{\partial x_j} = 2 \cdot e_i(x) \cdot \frac{\partial e_i(x)}{\partial x_j} \tag{20}$$

$$\frac{\partial \psi(x)}{\partial x_j} = \sum_{i=1}^{m} 2 \cdot e_i(x) \cdot \frac{\partial e_i(x)}{\partial x_j} \tag{21}$$

Where $\frac{\partial e_i(x)}{\partial x_j}$ is the element in the *i* column of the *j* row of the Jacobian matrix, thus the first-order derivative can also be expressed in the following form, as shown in Eq. 22.

$$\frac{\partial \psi_i(x)}{\partial x_j} = 2 \cdot J^T \cdot e(x) \tag{22}$$

Second-order derivative, as shown in Eq. 23.

$$\begin{aligned}\frac{\partial^2 \psi(x)}{\partial x_j \partial x_k} &= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{m} 2 \cdot e_i(x) \cdot \frac{\partial e_i(x)}{\partial x_j}\right) \\ &= 2\sum_{i=1}^{m}\left(\frac{\partial e_i(x)}{\partial x_j} \cdot \frac{\partial e_i(x)}{\partial x_k} + e_i(x) \cdot \frac{\partial^2 e_i(x)}{\partial x_j \partial x_k}\right)\end{aligned} \tag{23}$$

Observing the result of the second-order derivative, the terms $\frac{\partial e_i(x)}{\partial x_j}$ and $\frac{\partial e_i(x)}{\partial x_k}$ are elements of the Jacobian matrix. When the iterative point is far from the target point, both the error and its second-order derivative are small and can be ignored. Therefore, the second-order derivative can be expressed in the following form, as shown in Eq. 24.

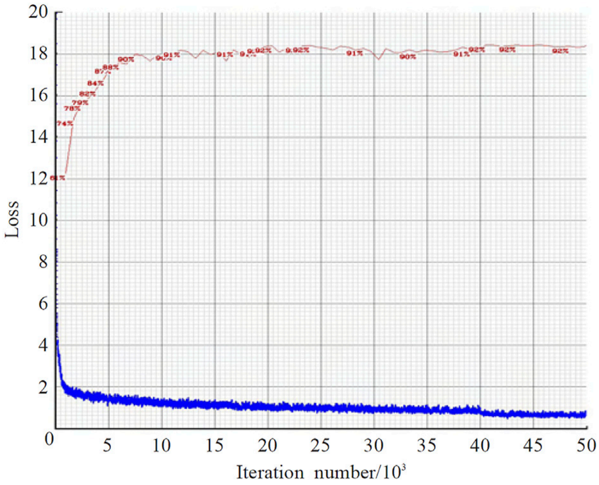


FIGURE 6
Model training loss convergence status.

$$\frac{\partial^2 \psi(x)}{\partial x_j \partial x_k} = 2 \cdot J^T \cdot J \tag{24}$$

Therefore, after the second-order expansion, $\psi(x)$ can be written in the following form, as shown in Eq. 25:

$$\psi(x) = \psi\left(x^{(k)}\right) + 2\left(x - x^{(k)}\right)^T J e(x) + \left(x - x^{(k)}\right)^T J^T J \left(x - x^{(k)}\right) \tag{25}$$

Similarly, by differentiating it and setting the derivative equal to zero, Eq. 26:

$$\nabla \psi(x) = 2J^T e\left(x^{(k)}\right) + 2J^T J\left(x - x^{(k)}\right) = 0 \tag{26}$$

Let $\triangle x = x - x^{(k)}$ then, as shown in Eq. 27:

$$\triangle x = -\left(J^T J\right)^{-1} \cdot J^T \cdot e \tag{27}$$

3.3 Vehicle speed measurement

Through prior estimation, $u_i(t)$ and $u_i(t - \Delta t)$ can be mapped to the world coordinate system, representing the actual distance moved by the target vehicle from the previous frame to the current frame, as shown in Eq. 28. $\|S_i\|$ is measured in meters and is the Euclidean norm of S_i, representing the physical distance moved by the target vehicle in the world coordinate system from time $t - \Delta t$ to t. The speed of the vehicle target can be measured using $\|S_i\|$ as Eq. 29. Here, Δt is the time between two frames, measured in seconds, and is considered constant, being the reciprocal of the frame rate. For highway surveillance videos, which typically have a frame rate of 25 fps, $\Delta t = 1/25$.

$$S_i = \varphi(a, b, c) \cdot u_i(t) - \varphi(a, b, c) \cdot u_i(t - \Delta t) \tag{28}$$

$$v_i = \frac{\|S_i\|}{\Delta t} = \frac{\|\varphi(a, b, c) \cdot u_i(t) - \varphi(a, b, c) \cdot u_i(t - \Delta t)\|}{\Delta t} \tag{29}$$

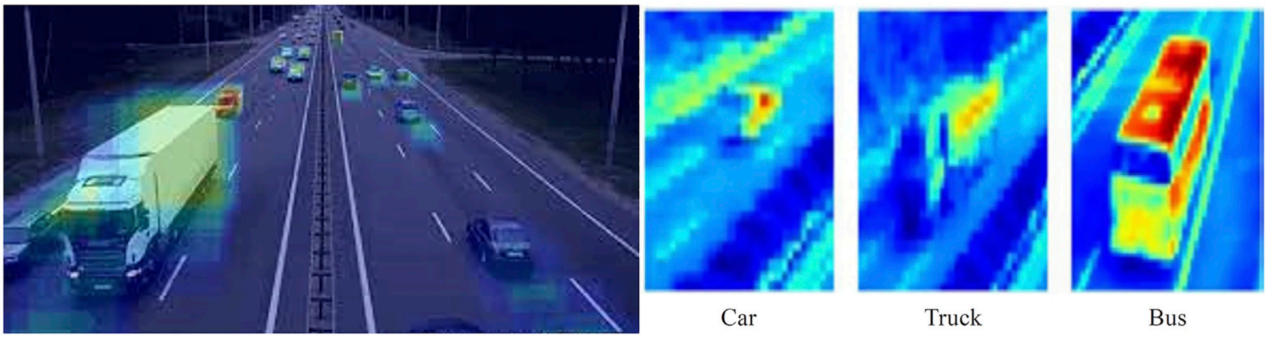


FIGURE 7
Classified target feature map.

Assuming a vehicle's trajectory contains m frame trajectory points, meaning in the first m frames of the video, the vehicle's speed between each adjacent pair of frames is $v1, v2, \ldots, v_{m-1}$, then according to Eq. 29, v1, v2, vm^{-1} as shown in Eqs 30–32:

$$\boldsymbol{v_1} = \frac{\|\boldsymbol{S_1}\|}{\Delta \boldsymbol{t}} = \frac{\|\boldsymbol{\varphi(a,b,c)\cdot u_2(t) - \varphi(a,b,c)\cdot u_1(t)}\|}{\Delta \boldsymbol{t}} \tag{30}$$

$$\boldsymbol{v_2} = \frac{\|\boldsymbol{S_2}\|}{\Delta \boldsymbol{t}} = \frac{\|\boldsymbol{\varphi(a,b,c)\cdot u_3(t) - \varphi(a,b,c)\cdot u_2(t)}\|}{\Delta \boldsymbol{t}} \tag{31}$$

$$\boldsymbol{v_{m-1}} = \frac{\|\boldsymbol{S_{m-1}}\|}{\Delta \boldsymbol{t}} = \frac{\|\boldsymbol{\varphi(a,b,c)\cdot u_m(t) - \varphi(a,b,c)\cdot u_{m-1}(t)}\|}{\Delta \boldsymbol{t}} \tag{32}$$

Therefore, the average driving speed of the target vehicle in the first m frames is as shown in Eq. 33. The detection of the target vehicle's speed is achieved by calculating the average of the instantaneous speeds over multiple frames.

$$\boldsymbol{v} = \frac{\sum_{i=1}^{m-1} \boldsymbol{v_i}}{\boldsymbol{m-1}} \tag{33}$$

4 Model training and evaluation metrics selection

4.1 Experimental environment and model training

Experimental setup and hardware environment for the dataset: System Type: Windows 10 64-bit Operating System, Memory: 64GB, GPU: NVIDIA GeForce RTX3080ti, 24 GB Graphics Card. Software environment: The auxiliary environment includes CUDA V11.2, OpenCV4.5.3. This article tested different corresponding datasets for various traffic scenarios. The dataset established in This article comprises a total of 30,000 images, including a diverse collection from different scenes, angles, and times.

During training, 80% of the dataset was used for training, while 20% of the data was reserved for testing. Data augmentation was applied in this study, which involved random scaling, cropping, and arrangement of images using the Mosaic method. Random rotation (parameter set to 0.5), random exposure (parameter set to 1.5), and saturation (parameter set to 1.5) were employed to enrich the training data. The learning rate was initially set to 0.001, and the maximum number of training iterations was set to 50,000. To optimize model convergence, the learning rate was adjusted to 0.0005 after 40,000 iterations. The input images to the network were resized to a resolution of 416 × 416, and a batch size of 8 was used during training to ensure efficient network processing. The convergence of the model's training loss and mAP (mean Average Precision) can be observed in Figure 6. It shows that the model converged around 3,000 iterations, and as the loss decreased, mAP also reached a high level.

Convolutional Neural Networks (CNNs) are capable of extracting key features from image objects. The detected objects are classified into three categories: Car, Truck, and Bus. The unique features of each class can be observed in Figure 7, where each class of object exhibits distinct characteristics within the convolutional network. These distinct features are used for classification and detection purposes.

4.2 Selection of evaluation metrics

To verify the effectiveness of the model's detection, several typical metrics in the field of object detection and classification were selected for evaluation. For distracted driving behavior detection and classification, the focus is on detection precision and recall rate, as well as classification accuracy. Therefore, the model is evaluated using precision, recall, and F1_Score.

AP (Average Precision) is the average accuracy and a mainstream evaluation metric for object detection models. To correctly understand AP, it is necessary to use three concepts: Precision, Recall, and *IoU* (Intersection over Union). *IoU* measures the degree of overlap between two areas, specifically the overlap rate between the target window generated by the model and the originally marked window, which represents the detection accuracy *IoU*. The calculation formula is shown in Eq. 34. In an ideal situation, *IoU* equals 1, indicating a perfect overlap.

$$\boldsymbol{IoU} = \frac{\boldsymbol{Detection\ Result} \cap \boldsymbol{Ground\ Truth}}{\boldsymbol{Detection\ Result} \cup \boldsymbol{Ground\ Truth}} \tag{34}$$

Precision and Recall in object detection: Assuming a set of images containing several targets for detection, Precision represents the proportion of targets detected by the model that are actual target objects, while Recall represents the proportion of all real targets detected by the model. TP (True Positive) denotes samples correctly identified as positive, TN (True Negative) denotes samples correctly identified as negative, FP (False Positive) denotes samples incorrectly identified as

TABLE 4 Fitting model results.

Number	Formulas	Abbreviation	AIC	BIC	R^2	p-value
1	$y = a^*x + b$	Line2p	139.69	141.6	0.774	3.3085e-05
2	$y = 1/(a^*x + b)$	Com2p	95.8	98.3	0.912	1.08e-08
3	$y = 1/(a^*x + b) + c$	Com3p	94.5	96.7	0.913	5.35e-07
4	$y = a^*\hat{x}2 + b^*x + c$	Line3p	118.0	121.0	0.958	2.59e-08
5	$y = a^*\, ln(x) + b$	Log2p	130.9	132.8	0.880	7.2456e-07
6	$y = a^*\, exp(b^*x)$	Exp2p	84.7	86.6	0.996	1.71e-15
7	$y = a^*\, exp(b^*x) + c$	Exp3p	**82.8**	**85.4**	**0.997**	**2.52e-14**

Bold values represent the method chosen in this article.

positive, and FN (False Negative) denotes samples incorrectly identified as negative. The calculation of Precision and Recall values relies on the formulas shown in Eqs 35, 36.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \tag{35}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \tag{36}$$

After calculating values using the formula, a PR (Precision-Recall) curve can be plotted. The AP (Average Precision) is the mean of Precision values on the PR curve. To achieve more accurate results, the PR curve is smoothed, and the area under the smoothed curve is calculated using integral methods to determine the final AP value. The calculation formula is shown as Eq. 37.

$$\boldsymbol{AP} = \int_0^1 \boldsymbol{P_{smooth}(r)dr} \tag{37}$$

The F1-Score, also known as the F1 measure, is a metric for classification problems, often used as the final metric in multi-class problems. It is the harmonic mean of precision and recall. For the F1-Score of a single category, the calculation formula is as shown in Eq. 38.

$$\boldsymbol{F1_k} = 2\frac{\boldsymbol{Recall_k \times Precision_k}}{\boldsymbol{Recall_k + Precision_k}} \tag{38}$$

Subsequently, calculate the average value for all categories, denoted as F1. The calculation formula is shown in Eq. 39.

$$\boldsymbol{F1} = \left(\frac{1}{n}\sum \boldsymbol{F1_k}\right)^2 \tag{39}$$

mAP (mean Average Precision) involves calculating the AP (Average Precision) for all categories and then computing the mean. The calculation formula is shown in Eq. 40.

$$\boldsymbol{mAP} = \frac{\sum \boldsymbol{AP_i}}{\boldsymbol{n}}, \boldsymbol{i = 1, 2, \ldots, n} \tag{40}$$

5 Results and discussion

5.1 Evaluation of object detection model results

Based on the aforementioned evaluation metrics, the trained object detection models are tested and assessed using the test sets from the datasets. The algorithm shows good statistical accuracy for different vehicle types, with APs of Car, Bus, Truck being 93.58, 91.26, 90.05 respectively, mAP at 92.42, and F1_Score at 97. This is primarily due to the high visibility in tunnel and roadbed sections, where target features are more distinct, resulting in a more accurate model. Overall, the model's detection accuracy for buses is lower than for other categories, mainly because the sample size for buses is significantly smaller than for other categories. However, with a mean Average Precision (mAP) exceeding 90%, it demonstrates that the proposed model is reliable and fully applicable to highway scenarios.

5.2 Evaluation of speed estimation results

5.2.1 Selection of optimal fitting model

Based on the data distribution, This article selects 7 video points for fitting analysis with 7 sets of linear and nonlinear data. This curve relationship is not intuitively obvious but requires statistical testing. The optimal fitting model is chosen by comparing the degree of fit and its significance. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two commonly used indicators for assessing model fitness, with smaller values indicating a better-fitting model. Therefore, before selecting a model, it is necessary to assess the AIC and BIC values for each model, including dependent and independent variables. Additionally, the goodness of fit R^2 and p-value are also key parameters for evaluating the quality of the fit. As the data distribution within the range of road video surveillance is essentially similar in terms of distance calibration, a random surveillance point is selected for the fitting analysis of the 7 formulas, with results as shown in Table 4.

From Table 4, it is evident that apart from linear fitting, the goodness of fit R^2 for all other methods is greater than 0.8. Among them, the $Exp3p$ fitting shows the best performance, hence $Exp3p$ is chosen as the formula for distance-speed fitting.

To obtain the best fitting parameters for $Exp3p$, employing Maximum Likelihood Estimation, Maximum A Posteriori Estimation, and Non-linear Least Squares method for parameter estimation on the distance calibration data from 7 video points. The parameters are evaluated using AIC, BIC, R^2, and p-value, with the evaluation results presented in Table 5; Figure 8.

From the above table, it is clear that for the $Exp3p$ parameter estimation of the 7 video points, Maximum Likelihood Estimation

TABLE 5 Parameter estimation results.

Number	MLE				MAP				NLS			
	AIC	BIC	R^2	*p*-value	AIC	BIC	R^2	*p*-value	AIC	BIC	R^2	*p*-value
1	**80**	**82**	**0.998**	**2.52e-14**	81	83	0.998	1.15e-14	82	85	0.997	2.52e-14
2	**83**	**86**	**0.994**	**1.76e-10**	84	87	0.994	4.32e-10	86	89	0.993	5.38e-10
3	**81**	**84**	**0.996**	**5.81e-13**	84	86	0.995	7.65e-13	84	87	0.995	8.26e-13
4	**94**	**100**	**0.923**	**5.63e-09**	96	103	0.913	4.25e-08	98	105	0.902	5.63e-08
5	**83**	**91**	**0.983**	**8.54e-13**	85	92	0.980	2.85e-12	85	93	0.975	3.16e-12
6	**84**	**87**	**0.992**	**2.52e-10**	85	88	0.995	5.15e-10	87	90	0.990	6.87e-10
7	**82**	**85**	**0.995**	**4.84e-12**	83	87	0.994	6.62e-12	86	89	0.993	5.36e-12

Bold values represent the method chosen in this article.

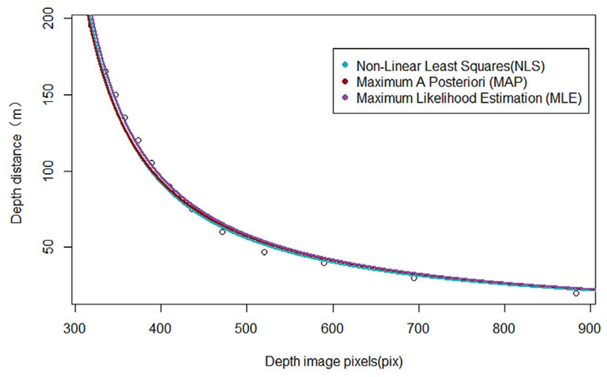


FIGURE 8
Parameter estimation results.

shows the best performance, followed by Maximum A Posteriori Estimation, and lastly Non-linear Least Squares method, as indicated by AIC, BIC, R^2, and *p*-value.

5.2.2 Speed estimation results

To evaluate the measurement results of the speed estimation method, based on radar and video multi-sensor fusion technology, the results measured by millimeter-wave radar are taken as the true speed values. The verification experiment was conducted in the Shimen Tunnel on the Hanping Expressway in Shaanxi China, where radar and video integration devices were installed at 150-m intervals, totaling seven units, to achieve holographic perception of traffic flow states within a 1050-m range, obtaining detailed information on coordinates, lane positions, and speeds for different lanes and vehicle types. Vehicle speeds detected by millimeter-wave radar and video were extracted using timestamps and target IDs. The comparison between the measured results and the true speed values, along with the overall experimental results and performance analysis, are shown in Table 6.

From Table 6, it is observed that the vehicle speed measurement method based on video, as discussed in This article, shows relatively good performance in scenarios with high overall speeds on highways. The minimum root mean square error is 2.0635, and the maximum is 9.2797. The main reasons for the larger deviation between the measured speeds and the actual values are environmental conditions, such as lighting and line shape. The coefficient of determination ranges from a minimum of 0.68259 to a maximum of 0.97730. The variation in the goodness of fit is for the same reasons as the minimum mean square error. Additionally, to further evaluate the speed tracking performance of this method, the vehicle speed measurement data from 7 video locations are manually divided into Front section, Middle section, Back section, and End section, for a comprehensive analysis of the overall tracking effect in these four segments, as seen in Figure 9.

As depicted in Figure 9, the effective measurement distance of this method is around 140 m, with the absolute speed error generally within 1–8 km/h, meeting the accuracy requirements for speed

TABLE 6 Overall speed measurement results and performance analysis.

Station number	MSE	RMSE	MAE	R^2
K733 + 953	26.9133	5.1878	3.8536	0.87993
K734 + 088	14.2012	3.7685	2.9179	0.90497
K734 + 843	52.2661	7.2295	5.3957	0.8889
K734 + 983	86.1127	9.2797	7.3639	0.68259
K735 + 123	6.6045	2.5699	2.0935	0.96168
K735 + 263	4.2581	2.0635	1.6984	0.97730
K735 + 403	81.6310	9.0205	7.6991	0.75010

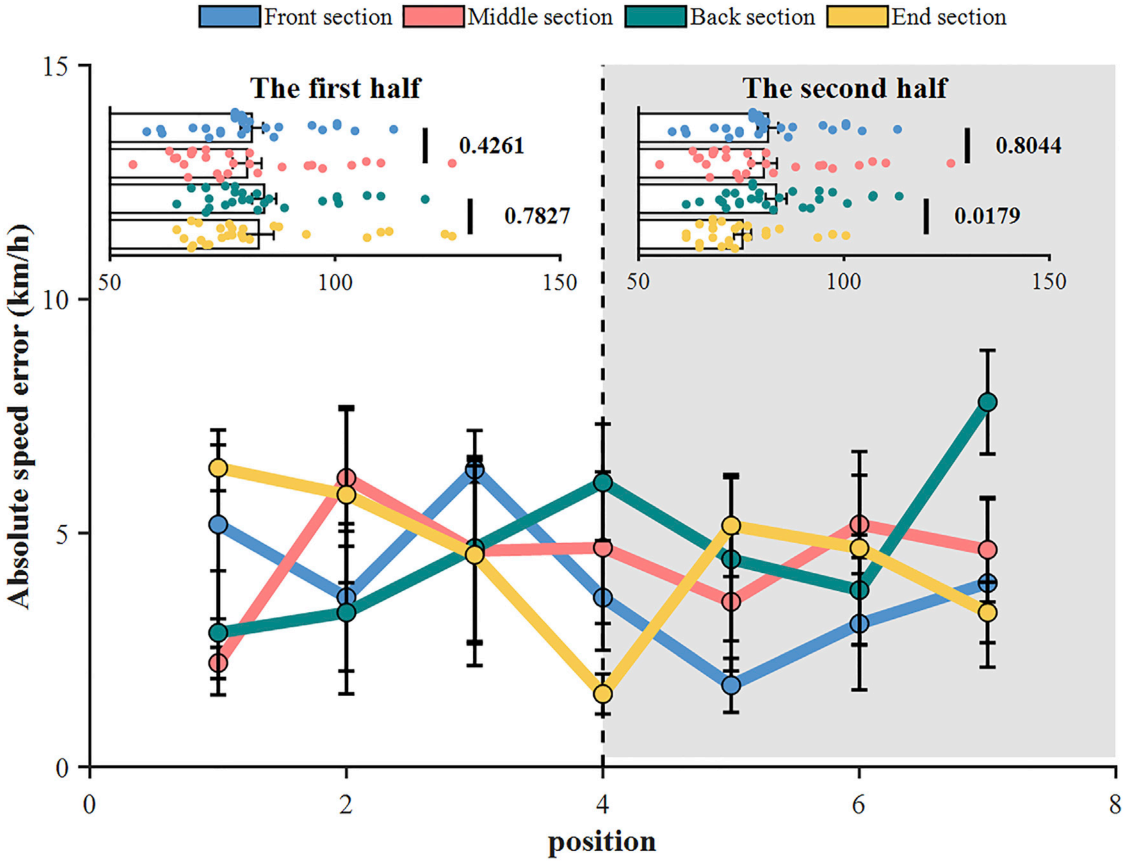


FIGURE 9
Analysis of speed tracking effect.

measurement. This method has certain advantages in distance detection, especially in tunnel scenarios, where a camera spacing of 150 m allows for continuous tracking of vehicle trajectories and speeds based on video. For further analysis of speed tracking differences within the 150 m detection range, it's divided into The first half and The second half. The first half data shows a minimum significance level of 0.4261, indicating small differences in speed tracking, reflecting stable tracking performance. The second half data has a minimum significance level of 0.0179, indicating some fluctuations in speed in the End section of The second half, but the absolute speed error still shows good precision.

6 Conclusion

This article proposes an improved YOLOv5s + DeepSORT vehicle speed measurement algorithm for surveillance videos in highway scenarios, capable of vehicle target detection and continuous speed tracking without camera prior parameters and calibration. The main conclusions are as follows:

(1) The introduction of the Swin Transformer Block module improves the model's ability to capture local areas of interest, effectively increasing the detector's accuracy; using *CIoU* Loss to replace the original *GIoU* loss further enhances the detector's localization precision and effectively reduces omissions in congested vehicle scenarios; the algorithm shows good statistical accuracy for different vehicle types, with APs of Car, Bus, Truck being 93.58, 91.26, 90.05 respectively, mAP at 92.42, and F1_Score at 97.

(2) A calibration algorithm for traffic monitoring scenarios was proposed, which uses known reference points such as the image's centerline and contour marks. It applies Maximum Likelihood Estimation, Maximum A Posteriori Estimation, and Non-linear Least Squares method for the conversion between image pixel coordinates and actual coordinates. The parameter estimation showed good results, with Maximum Likelihood Estimation being the best, and AIC, BIC, R^2, and p-value being 83.56, 87.86, and 8.66E-10 respectively.

(3) The vehicle speed measurement is achieved by calculating the average of instantaneous speeds over multiple frames. This method's effective measurement distance is about 140m, with an absolute speed error generally within 1–8 km/h, meeting the accuracy requirements for speed measurement. It has certain advantages in distance detection, especially in tunnel scenarios where a camera spacing of 150 m allows for continuous tracking of vehicle trajectories and speeds based on video.

(4) However, during experiments, it was found that vehicle speed accuracy is influenced by road geometry, environmental conditions, lighting, resolution, etc., These can be mitigated

through image enhancement optimization algorithms or by increasing video resolution, thus achieving more accurate vehicle speed measurements, which help regulatory bodies more effectively control speeds on the roads, reducing instances of speeding and thereby decreasing traffic accidents, enhancing road safety. Additionally, with the rapid development of multi-sensor fusion technology, the integration of video and millimeter-wave radar detection results can complement each other, providing technical support for active traffic safety management on highways.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ZL: Conceptualization, Methodology, Writing–original draft. YB: Funding acquisition, Writing–original draft. XY: Project administration, Resources, Writing–review and editing. YL: Investigation, Writing–review and editing. SY: Software, Writing–review and editing. MW: Funding acquisition, Project administration, Writing–review and editing. QY: Data curation, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China (52208424); the Natural Science Foundation of Chongqing (2022NSCQ-MSX1939); Chongqing Municipal Education Commission Foundation (KJQN202300728); Chongqing Talent Innovation Leading Talent Project (CQYC20210301505); The Key Research and Development Program of Guangxi, China, (Grant No. AB21196034).

Acknowledgments

The authors would like to thank the support of their colleagues in the Research and Development Center of Transport Industry of Self-driving Technology.

Conflict of interest

Authors ZL, XY, SY, MW, and QY were employed by China Merchants Chongqing Communications Research and Design Institute Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Wan Y, Huang Y, Buckles B. Camera calibration and vehicle tracking: highway traffic video analytics. *Transp Res C Emerg Technol* (2014) 44:202–13. doi:10.1016/j.trc.2014.02.018

2. Karoń G, Mikulski J. Selected problems of transport modelling with ITS services impact on travel behavior of users. In: 2017 15th International Conference on ITS Telecommunications (ITST); 29-31 May 2017; Warsaw, Poland. IEEE (2017). p. 1–7. doi:10.1109/ITST.2017.7972231

3. Wang Y, Yu C, Hou J, Chu S, Zhang Y, Zhu Y. ARIMA model and few-shot learning for vehicle speed time series analysis and prediction. *Comput Intell Neurosci* (2022) 2022:1–9. doi:10.1155/2022/2526821

4. Jia S, Peng H, Liu S. Urban traffic state estimation considering resident travel characteristics and road network capacity. *J Transportation Syst Eng Inf Tech* (2011) 11: 81–5. doi:10.1016/S1570-6672(10)60142-0

5. Javadi S, Dahl M, Pettersson MI. Vehicle speed measurement model for video-based systems. *Comput Electr Eng* (2019) 76:238–48. doi:10.1016/j.compeleceng.2019.04.001

6. Dahl M, Javadi S. Analytical modeling for a video-based vehicle speed measurement framework. *Sensors (Switzerland)* (2020) 20(1):160. doi:10.3390/s20010160

7. Khan A, Sarker DMSZ, Rayamajhi S. Speed estimation of vehicle in intelligent traffic surveillance system using video image processing. *Int J Sci Eng Res* (2014) 5(12): 1384–90. doi:10.14299/ijser.2014.12.003

8. Wicaksono DW, Setiyono B. Speed estimation on moving vehicle based on digital image processing. *Int J Comput Sci Appl Math* (2017) 3(1):21–6. doi:10.12962/j24775401.v3i1.2117

9. Lu S, Wang Y, Song H. A high accurate vehicle speed estimation method. *Soft Comput* (2020) 24:1283–91. doi:10.1007/s00500-019-03965-w

10. Liu C, Huynh DQ, Sun Y, Reynolds M, Atkinson S. A vision-based pipeline for vehicle counting, speed estimation, and classification. *IEEE Trans Intell Transportation Syst* (2021) 22:7547–60. doi:10.1109/TITS.2020.3004066

11. Bhardwaj R, Tummala GK, Ramalingam G, Ramjee R, Sinha P. AutoCalib: automatic traffic camera calibration at scale. *ACM Trans Sen Netw* (2018) 14(3-4):1–27. doi:10.1145/3199667

12. Qimin X, Xu L, Mingming W, Bin L, Xianghui S. A methodology of vehicle speed estimation based on optical flow. In: Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics; 08-10 October 2014; Qingdao, China. IEEE (2014). p. 33–7. doi:10.1109/SOLI.2014.6960689

13. Schoepflin TN, Dailey DJ. Dynamic camera calibration of roadside traffic management cameras for vehicle speed estimation. *IEEE Trans Intell Transportation Syst* (2003) 4:90–8. doi:10.1109/TITS.2003.821213

14. Han J, Heo O, Park M, Kee S, Sunwoo M. Vehicle distance estimation using a mono-camera for FCW/AEB systems. *Int J Automotive Tech* (2016) 17:483–91. doi:10.1007/s12239-016-0050-9

15. Sochor J, Juranek R, Spanhel J, Marsik L, Siroky A, Herout A, et al. Comprehensive data set for automatic single camera visual speed measurement. *IEEE Trans Intell Transportation Syst* (2019) 20:1633–43. doi:10.1109/TITS.2018.2825609

16. Lin H-Y, Li K-J, Chang C-H. Vehicle speed detection from a single motion blurred image. *Image Vis Comput* (2008) 26:1327–37. doi:10.1016/j.imavis.2007.04.004

17. Celik T, Kusetogullari H. Solar-powered automated road surveillance system for speed violation detection. *IEEE Trans Ind Elect* (2010) 57:3216–27. doi:10.1109/TIE.2009.2038395

18. Nguyen TT, Pham XD, Song JH, Jin S, Kim D, Jeon JW. Compensating background for noise due to camera vibration in uncalibrated-camera-based vehicle

speed measurement system. *IEEE Trans Veh Technol* (2011) 60:30–43. doi:10.1109/TVT.2010.2096832

19. Eslami H, Raie AA, Faez K. Precise vehicle speed measurement based on a hierarchical homographic transform estimation for law enforcement applications. *IEICE Trans Inf Syst* (2016) E99.D:1635–44. doi:10.1587/transinf.2015EDP7371

20. Famouri M, Azimifar Z, Wong A. A novel motion plane-based approach to vehicle speed estimation. *IEEE Trans Intell Transportation Syst* (2019) 20:1237–46. doi:10.1109/TITS.2018.2847224

21. Li J, Chen S, Zhang F, Li E, Yang T, Lu Z. An adaptive framework for multi-vehicle ground speed estimation in airborne videos. *Remote Sens (Basel)* (2019) 11:1241. doi:10.3390/rs11101241

22. Koyuncu H, Koyuncu B. Vehicle Speed detection by using Camera and image processing software. *Int J Eng Sci (Ghaziabad)* (2018) 7:64–72. doi:10.9790/1813-0709036472

23. Kim J-H, Oh W-T, Choi J-H, Park J-C. Reliability verification of vehicle speed estimate method in forensic videos. *Forensic Sci Int* (2018) 287:195–206. doi:10.1016/j.forsciint.2018.04.002

24. Sochor J, Juránek R, Herout A. Traffic surveillance camera calibration by 3D model bounding box alignment for accurate vehicle speed measurement. *Computer Vis Image Understanding* (2017) 161:87–98. doi:10.1016/j.cviu.2017.05.015

25. Palubinskas G, Kurz F, Reinartz P. Model based traffic congestion detection in optical remote sensing imagery. *Eur Transport Res Rev* (2010) 2:85–92. doi:10.1007/s12544-010-0028-z

26. Doğan S, Temiz MS, Külür S. Real time speed estimation of moving vehicles from side view images from an uncalibrated video camera. *Sensors* (2010) 10(5):4805–24. doi:10.3390/s100504805

27. Li S, Yu H, Zhang J, Yang K, Bin R. Video-based traffic data collection system for multiple vehicle types. *IET Intell Transport Syst* (2014) 8:164–74. doi:10.1049/iet-its.2012.0099

28. Jeyabharathi D, Dejey DD. Vehicle tracking and speed measurement system (VTSM) based on novel feature descriptor: diagonal hexadecimal pattern (DHP). *J Vis Commun Image Represent* (2016) 40:816–30. doi:10.1016/j.jvcir.2016.08.011

29. Agrawal SC, Tripathi RK. An image processing based method for vehicle speed estimation. *Int J Scientific Tech Res* (2020) 9:1241–6.

30. Biswas D, Su H, Wang C, Stevanovic A. Speed estimation of multiple moving objects from a moving UAV platform. *ISPRS Int J Geoinf* (2019) 8(6):259. doi:10.3390/ijgi8060259

31. Lee J, Roh S, Shin J, Sohn K. Image-based learning to measure the space mean speed on a stretch of road without the need to tag images with labels. *Sensors (Switzerland)* (2019) 19:1227. doi:10.3390/s19051227

32. Dong H, Wen M, Yang Z. Vehicle speed estimation based on 3D ConvNets and non-local blocks. *Future Internet* (2019) 11(6):123. doi:10.3390/fi11060123

33. Luvizon DC, Nassu BT, Minetto R. A video-based system for vehicle speed measurement in urban roadways. *IEEE Trans Intell Transportation Syst* (2017) 18:1–12. doi:10.1109/TITS.2016.2606369

34. Yang L, Li M, Song X, Xiong Z, Hou C, Qu B. Vehicle speed measurement based on binocular stereovision system. *IEEE Access* (2019) 7:106628–41. doi:10.1109/ACCESS.2019.2932120

35. Blankenship K, Diamantas S. Detection, tracking, and speed estimation of vehicles: a homography-based approach. *IMPROVE* (2022) 1:211–8. doi:10.5220/0011093600003209

36. Fernández Llorca D, Hernández Martínez A, García Daza I. Vision-based vehicle speed estimation: a survey. *IET Intell Transport Syst* (2021) 15:987–1005. doi:10.1049/itr2.12079

37. Kim HJ. Vehicle detection and speed estimation for automated traffic surveillance systems at nighttime. *Tehnicki Vjesnik* (2019) 26:091448. doi:10.17559/TV-20170827091448

38. Ashraf MH, Jabeen F, Alghamdi H, Zia MS, Almutairi M. HVD-net: a hybrid vehicle detection network for vision-based vehicle tracking and speed estimation. *J King Saud Univ - Comp Inf Sci* (2023) 35:101657. doi:10.1016/j.jksuci.2023.101657

39. Pal SK, Pramanik A, Maiti J, Mitra P. Deep learning in multi-object detection and tracking: state of the art. *Appl Intelligence* (2021) 51:6400–29. doi:10.1007/s10489-021-02293-7

40. Jiao L, Zhang F, Liu F, Yang S, Li L, Feng Z, et al. A survey of deep learning-based object detection. *IEEE Access* (2019) 7:128837–68. doi:10.1109/ACCESS.2019.2939201

41. Khosravi H, Dehkordi RA, Ahmadyfard A. Vehicle speed and dimensions estimation using on-road cameras by identifying popular vehicles. *Scientia Iranica* (2022) 29. doi:10.24200/sci.2020.55331.4174

42. Huang L, Zhe T, Wu J, Wu Q, Pei C, Chen D. Robust inter-vehicle distance estimation method based on monocular vision. *IEEE Access* (2019) 7:46059–70. doi:10.1109/ACCESS.2019.2907984

43. Jamshidnejad A, De Schutter B. Estimation of the generalised average traffic speed based on microscopic measurements. *Transportmetrica A: Transport Sci* (2015) 11:525–46. doi:10.1080/23249935.2015.1026957

44. Sarkar NC, Bhaskar A, Zheng Z, Miska MP. Microscopic modelling of area-based heterogeneous traffic flow: area selection and vehicle movement. *Transp Res Part C Emerg Technol* (2020) 111:373–96. doi:10.1016/j.trc.2019.12.013

45. Appathurai A, Sundarasekar R, Raja C, Alex EJ, Palagan CA, Nithya A. An efficient optimal neural network-based moving vehicle detection in traffic video surveillance system. *Circuits Syst Signal Process* (2020) 39:734–56. doi:10.1007/s00034-019-01224-9

46. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. *Proc IEEE Int Conf Comp Vis* (2021) 10012–22. doi:10.1109/ICCV48922.2021.00986

47. Zhang Q, Zhang M, Chen T, Sun Z, Ma Y, Yu B. Recent advances in convolutional neural network acceleration. *Neurocomputing* (2019) 323:37–51. doi:10.1016/j.neucom.2018.09.038

48. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: Proceedings - International Conference on Image Processing, ICIP; 17-20 September 2017; Beijing, China. IEEE (2016). p. 3464–8. doi:10.1109/ICIP.2016.7533003

49. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: Proceedings - International Conference on Image Processing, ICIP; 17-20 September 2017; Beijing, China. IEEE (2017). p. 3645–9. doi:10.1109/ICIP.2017.8296962

50. Liu X, Liu W, Mei T, Ma H. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: *Lecture notes in computer science including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics*. Berlin, Germany: Spinger (2016). p. 869–84. doi:10.1007/978-3-319-46475-6_53

TYPE Original Research
PUBLISHED 07 March 2024
DOI 10.3389/fphy.2024.1361695

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Guodong Du,
Yanshan University, China
Peng Gui,
Wuhan University, China

*CORRESPONDENCE
Yan Xiang,
sharonxiang@126.com

RECEIVED 26 December 2023
ACCEPTED 20 February 2024
PUBLISHED 07 March 2024

CITATION
Xian Y, Qin Y and Xiang Y (2024), Auto-verbalizer filtering for prompt-based aspect category detection.
Front. Phys. 12:1361695.
doi: 10.3389/fphy.2024.1361695

Auto-verbalizer filtering for prompt-based aspect category detection

Yantuan Xian[1,2], Yuan Qin[1,2] and Yan Xiang[1,2]*

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [2]Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, China

Aspect category detection (ACD) is a basic task in sentiment analysis that aims to identify the specific aspect categories discussed in reviews. In the case of limited label resources, prompt-based models have shown promise in few-shot ACD. However, their current limitations lie in the manual selection or reliance on external knowledge for obtaining the verbalizer, a critical component of prompt learning that maps predicted words to final categories. To solve these issues, we propose an ACD method to automatically build the verbalizer in prompt learning. Our approach leverages the semantic expansion of category labels as prompts to automatically acquire initial verbalizer tokens. Additionally, we introduce an indicator mechanism for auto-verbalizer filtering to obtain reasonable verbalizer words and improve the predicting aspect category reliability of the method. In zero-shot task, our model exhibits an average performance improvement of 7.5% over the second-best model across four ACD datasets. For the other three few-shot tasks, the average performance improvement over the second-best model is approximately 2%. Notably, our method demonstrates effectiveness, particularly in handling general or miscellaneous category aspects.

KEYWORDS

aspect category detection, prompt learning, few-shot, verbalizer, sentiment analysis

1 Introduction

Aspect category detection (ACD) is a subtask of sentiment analysis that aims to detect the categories contained in reviews from a predefined set of aspect categories. For example, the sentence "Nevertheless the food itself is pretty good" contains the aspect category "Food," and the sentence "But the staff was so horrible to us" contains the aspect category "Service." Most of the existing excellent methods [1–3] finetune the pre-trained language models to solve ACD tasks, and their effects largely depend on the size of labeled data. However, as online reviews are updated quickly, the aspect categories will also be updated. It is difficult to provide sufficient label data for newly emerging categories. Therefore, the performance of the above methods will drop significantly when there are only few labeled samples.

In order to stimulate pre-trained language models (PLMs) to exhibit a greater performance under the conditions of few-shot and zero-shot, the researchers were inspired by GPT-3 [4] and LAMA [5] and proposed to use prompt to convert the classification task into a cloze task, which unified the downstream task and PLMs into the same schema to maximize the use of prior knowledge of PLMs. Prompt learning obtains the probability of each token filled in the [MASK] position in the PLM vocabulary through the prompt and then uses the verbalizer to map it to the final category. As one of the

TABLE 1 Examples of verbalizer words in the “miscellaneous” category. Bold indicates that it appears in the other categories, and “xx” indicates that it does not appear in the PLM vocabulary.

Method	Prompt words
Manual	Miscellaneous,. . .
Search based	Bryan, anonymous, Wes, noise, LM, KH, Ethan, Wayne, dark, iii, YOU. . .
KBs	Sundry, assorted, heterogeneous, multifarious, extraneous, mixed. . .

important components of prompt learning, the verbalizer contains the mapping relationship between prompt words and the final aspect category. Therefore, constructing a high-quality prompt word set can greatly improve the performance of the verbalizer.

The current methods of constructing prompt word sets can be roughly divided into three types: manual construction [6], search based [7], and continuously learnable [8]. Table 1 shows the prompt words selected for the “miscellaneous” category of the restaurant dataset by different methods. It can be seen that the main problems are as follows: 1) These methods either require intensive manual work or require the support of external knowledge bases and labeled data and, thus, cannot handle zero-shot tasks at a small cost. At the same time, many words searched from external knowledge will not appear in the PLM vocabulary. For example, for the words highlighted in red in the third row of Table 1, statistics show that 11 of the first 50 prompt words obtained for this category cannot be recognized by PLMs. This is because the vocabulary of the external knowledge base is different from the PLMs, due to which the overlap between the two is lacking. 2) The manually constructed prompt word sets only contain category words themselves, so the diversity of prompt words is not enough. For example, the first row in Table 1 only contains the word “miscellaneous” itself. Search-based methods do not consider the specificity of prompt words, where a word may appear in different word sets. For example, words such as “anonymous” in the second row of Table 1 also appear in prompt word sets of other categories at the same time. In addition, these methods do not consider the characteristics of the ACD task, such as categories are basically represented by nouns.

In response to the first type of problems mentioned above, we propose to use the semantic expansions of category labels as prompts to directly search for the initial prompt words from the internal vocabulary of PLMs so that the prompt words in the verbalizer conform to the PLM vocabulary. For the second type of problem, we propose a filtering mechanism to select prompt words. Specifically, we first consider the task characteristics; that is, the ACD task is to detect predefined aspect categories contained in sentences which should be represented by words with actual meaning. Therefore, we start from the parts of speech and select nouns, verbs, and adjectives. Second, we consider diversity and select words with high semantic similarity to the category. Finally, in terms of specificity, choosing words that are much more similar to the category to which they belong than to other categories as prompt words can avoid confusion in the mapping process. The main contributions of this article are as follows:

1) Auto-verbalizer filtering methods are proposed for prompt-based aspect category detection, which alleviates the limitations of the detection performance caused by unreasonable verbalizer design in existing prompt-based ACD methods.
2) The semantic extension of category labels is used as prompts to construct an initial verbalizer and eliminate dependence on labeled data and external knowledge bases. At the same time, an automatic filtering mechanism is introduced for the verbalizer to select prompt words related to aspect categories.
3) Experiments show that the proposed method can achieve optimal performance under zero-shot and few-shot conditions compared with existing prompt-based learning methods.

2 Related work

2.1 Aspect category detection

Semeval proposed the ACD subtask in 2014. Under the condition of sufficient labeled data, most of the previous ACD methods are based on machine learning, such as the classic SVM [9] and maximum entropy [7,10] which handcrafts multiple features such as n-grams and lexical features to train a set of classification devices. In recent years, methods based on deep neural networks [2] have been widely adopted. In [11], the output of CNN training as a type of feature and other POS tags and other features was sent into the one-vs-all classifier. The one-vs-all classifier used in [3] consists of a set of CNN network layers above the LSTM layer, which implements aspect category detection and aspect term extraction in parallel.

2.2 Prompt verbalizer construction

In the case of insufficient labeled data, researchers detect categories by mining association rules [12] or calculating word co-occurrence frequencies [13], but this requires obtaining reasonable rules in advance. Since the release of GPT-3, prompt learning has provided new ideas for ACD when labels are insufficient. The way of using prompts to stimulate internal knowledge of PLMs and avoiding the introduction of a large number of parameters to be trained usually includes two important parts: templates and a verbalizer. According to the manually created cloze template provided by the LAMA dataset, the previous templates are all artificially created auxiliary sentences which are human-understandable. For example, manually designed prefix-type prompts [4] had achieved good results in some NLP tasks, such as text question answering and neural machine translation. However, although this type of template has the advantage of being intuitive, it requires a lot of experience and a lot of time to obtain good performance prompts and cannot be optimized to the best. To solve these problems, automated template-based methods are proposed [14–17], which automatically search for natural language phrases in discrete space to form prompt templates. Later, scholars discovered that the prompts were constructed to allow PLMs to better understand the task rather than humans. Therefore, they proposed that templates do not need to be limited to human understandability. In [18–21], continuous templates were directly constructed in the model embedding space. The template is no longer restricted by additional parameters and can itself be trained and optimized along with downstream tasks.

When working on templates, researchers are also focusing on exploring another important component of prompt learning—the construction of the verbalizer. The most straightforward method is to use manually selected words to construct prompt word sets, and it has been proven to be effective [7]. However, this type of method involves personal biases, so the coverage of the vocabulary is insufficient. Based on these problems, some automatically search-based methods have been proposed. The work in [22] searched for label words in the pruned candidate space and redefined the k classification problem as a binary classification problem of "1 vs. k-1" so that PLMs can distinguish category y from other categories. In [21], a two-stage gradient-based automatic search method was used to calculate the representation of each category in the first stage and train a classifier. The second stage uses this classifier to select words that are close to the category representation to construct a verbalizer. In [23–25], relevant words were selected from the external knowledge base and then refined to align with the PLM vocabulary. However, such automatically search-based methods require the assistance of sufficient training data or external knowledge. In contrast to the discrete verbalizer, the continuous verbalizer [8,20] represents categories in word embedding space and can be trained and optimized. In [8], vector form was used to represent categories, carry out a dot product between the token vector predicted by PLMs and the category vector, and select the corresponding category that obtains the maximum dot product as the prediction result. In [26], the filled-in token vectors of all sentences under each category were averaged to obtain the prototype representation of this category, and this prototype was continuously optimized. Similarly, continuous vectors also require a large amount of data for training and optimization, so they cannot be directly applied to zero-shot learning.

3 Prompt-based aspect category detection with auto-verbalizer filtering

3.1 Task definition

ACD is to identify aspect categories $y \in \{1, 2, \ldots C\}$ for a given sentence, where C is the number of aspect categories. The basic process of prompt-based ACD is as formula (1): the ith sentence x_i is packed into x_i^p with a template, which is a natural language text with the "[MASK]" token:

$$x_i^p\text{: } x_i \text{ [sep] It is about [MASK] category.} \qquad (1)$$

We obtain the probability p ([MASK]= $v|x_i^p$) of each token v in the vocabulary $V \in R^D$ filling in the [MASK] position by PLMs. The probability distribution vector of the entire vocabulary for the ith sentence is $P_i^V \in R^{1\times D}$. Finally, the probability of category y can be calculated as formula (2)

$$p\left([\text{MASK}] = v|x_i^p\right) = f\left(p\left([\text{MASK}] = v|x_i^p\right)|v \in V_p\right), \qquad (2)$$

where V_p is the prompt word set of the verbalizer and f is a function transforming the probability of prompt words into the probability of the category.

3.2 Initial construction of the verbalizer based on label semantic extension

When evaluating aspect categories of reviews, the most important consideration is the semantic similarity between the review and the label categories [27–29]. Consequently, the specific category label itself serves as valuable prior knowledge that can be utilized. Following this idea, we propose to utilize category labels as prompts to construct the verbalizer.

Specifically, as shown in Figure 1A, we use task-specific templates such as "[x]. This is about the [MASK] category," where [x] is the definition statement of the corresponding category label j in Wikipedia (see Table 2). The definition statement is encapsulated into a natural language text x_j^c with [MASK] tokens and is sent to PLMs to obtain the probability that each token in the vocabulary V is filled to the [MASK] position. In this way, the probability distribution vector $P_j^V \in R^{1\times D}$ for a given label category j can be obtained. As shown in Figure 1B, this is carried out for different label categories, and a complete verbalizer initial probability matrix $P \in R^{C\times D}$ is constructed.

3.3 Indicator mechanism for verbalizer filtering

We propose an indicator-based filtering mechanism to improve the verbalizer. Specifically, we set an indicator value b_{ji} for each probability p_{ji} in the probability matrix P representing the correlation of token i with a specific category j. A value of 1 signifies that the token is highly important for the corresponding category, whereas a value of 0 signifies the opposite. Initially, all indicator values are set to 1, forming the indicator matrix $B \in R^{C\times D}$. Next, as shown in Figure 1C, we refine the indicator matrix to obtain more reasonable prompt words by considering three parts.

(1) In order to be more consistent with the characteristics of the ACD task, we use the pos_tag method from the nltk package to define the set of tokens in the vocabulary V that match nouns, verbs, and adjectives as $\{pos\}$ and then adjust the corresponding element values in the indicator matrix B to get a new indicator matrix B^{pos} according to formula (3):

$$b_{ji}^{pos} = \begin{cases} b_{ji} & if\ v_i \in \{pos\} \\ 0 & else \end{cases} \qquad (3)$$

(2) In order to retain prompt words with more highly semantic similarity to a specific category, we further modify the element values in the matrix B^{pos} based on category semantic relevance and obtain B^{sem} according to formula (4):

$$b_{ji}^{sem} = \begin{cases} b_{ji}^{pos} & if\ p_{ji} > MAX\_M\left(P_j^V\right) \\ 0 & else \end{cases}, \qquad (4)$$

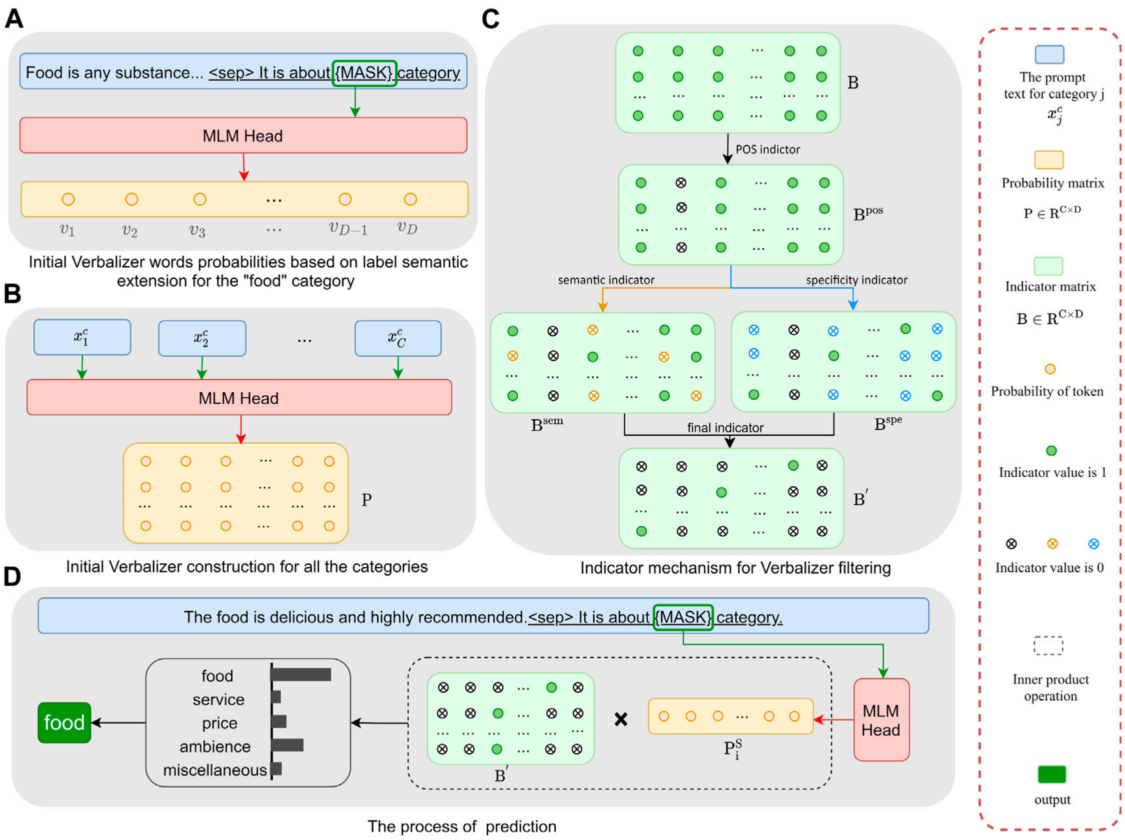

A
Food is any substance... <sep> It is about {MASK} category
MLM Head
v_1 v_2 v_3 ... v_{D-1} v_D
Initial Verbalizer words probabilities based on label semantic extension for the "food" category
B
x_1^c x_2^c ... x_C^c
MLM Head
P
Initial Verbalizer construction for all the categories
C
B
POS indictor
B^{pos}
semantic indicator
specificity indicator
B^{sem}
final indicator
B^{spe}
B'
Indicator mechanism for Verbalizer filtering
D
The food is delicious and highly recommended.<sep> It is about {MASK} category.
food
service
price
ambience
miscellaneous
food
B'
P_i^S
MLM Head
The process of prediction
The prompt text for category j x_j^c
Probability matrix $P \in R^{C \times D}$
Indicator matrix $B \in R^{C \times D}$
Probability of token
Indicator value is 1
Indicator value is 0
Inner product operation
output

FIGURE 1
Illustration of our model. **(A)** Initial Verbalizer words probabilities based on label semantic extension for the "food" category. **(B)** Initial Verbalizer construction for all the categories. **(C)** Indicator mechanism for Verbalizer filtering. **(D)** The process of prediction.

TABLE 2 Semantic extensions of categories. The semantic extensions are derived from Wikipedia or Baidu Encyclopedia. We take the first one or two sentences of the definition as the semantic extensions.

Label	Semantic extension
Food	Food is any substance consumed by an organism for nutritional support
Service	Customer service refers to the provision of assistance to customers or clients
Price	Price is the quantity of payment or compensation given by one party to another in return for goods or services
Ambience	Ambience which is also known as atmospheres or backgrounds
Miscellaneous	Miscellaneous refers to a collection of writings on various subjects or topics
Comfort	Comfort is the physical and psychological sense of ease
Size	Clothing size in general is the magnitude or dimensions of a thing
Quality	Quality is a product or service free of deficiencies
Layout	Keyboard layout is an arrangement of the keys on a typographic keyboard
Connection	Connection refers to a communication link between two or more devices
Service	Customer service is the assistance and advice provided by a company to those people who buy or use its products or services
Image	A digital image is an image composed of picture elements which is also known as pixels
Sound	The sound is the loudness of the sound and the characteristics of the timbre

where $MAX\_M(.)$ represents the Mth largest probability value in the probability distribution vector of the label category.

(3) In order to select the prompt words with specificity, we adjust the element values in the matrix B^{pos} based on the following formula to obtain the updated indicator matrix B^{spe} according to formula (5):

$$b_{ji}^{spe} = \begin{cases} b_{ji}^{pos} & if \dfrac{p_{ji}}{\sum_{j=1}^{C} p_{ji}} > \alpha \\ 0 & else \end{cases}, \tag{5}$$

where α is a threshold indicating that the words exceeding this threshold are class-specific.

Also, the modified matrix B' is calculated as formula (6)

$$B' = B^{sem} \circ B^{spe}, \tag{6}$$

where $\circ$ represents the Hadamard product.

Finally, the prompt words of each category are composed of tokens whose indicator value is 1 in the matrix B' under this category.

3.4 Aspect category prediction

During category prediction, we package the review x_i into a natural language text like in Figure 1D and send it to PLMs to obtain a probability distribution vector $P_i^S \in R^{1\times D}$ and finally map it to the aspect category label by the constructed verbalizer.

For the zero-shot scenario, we assume that all prompt words in the verbalizer contribute equally to the prediction of the corresponding category, so we calculate the category probability $\widehat{Y_{ij}}$ of the sentence x_i with respect to category j using the following formula (7):

$$\widehat{Y_{ij}} = P_i^S \left(B_j'\right)^T. \tag{7}$$

For few-shot scenario, we set a weight parameter for each token, and the probability $\widehat{Y_{ij}}$ of the sentence x_i with respect to category j is calculated as formula (8)

$$\widehat{Y_{ij}} = \left(P_i^S \circ W\right)\left(B_j'\right)^T, \tag{8}$$

where $W \in R^{1\times D}$ is the parameter vector to be trained, which can be optimized using the cross-entropy loss as formula (9). The objective function is the loss between the final predicted label and the true label:

$$loss = -\frac{1}{C} \sum_{i\in|D_{train}|} \sum_{j\in C} \hat{y} log p\left(y_j|x_i\right), \tag{9}$$

where $\hat{y}$ is the true label of input x_i.

4 Experiment

4.1 Datasets

We conducted experiments on four ACD datasets, including Restaurant-2014, Boots, Keyboards, and TV of the Amazon dataset. In the few-shot experiment, following most few-shot learning settings, we adopt the N way K shot mode, randomly selecting K samples of each category for the validation set and the training set, and the remaining samples are used as the test set. The size of the training set and validation set are $|D_{dev}|$=$|D_{train}|$=N * K.

4.2 Baselines

We selected several advanced models for comparative experiments. Same as this model, all prompt learning methods adopt the most basic prompt learning method: templates were used to convert the input into a natural language text with the [MASK] token, and the vocabulary token probability output by the model is mapped to class labels by the verbalizer. All models use the same template, so only the verbalizer is constructed differently.

Finetuning: The traditional finetuning methods add a classification layer after the PLM model, obtaining the hidden vector of [CLS] and making predictions via the classification layer.

Manual: The manually constructed verbalizer contains limited category prompt words. In this experiment, we use the category word itself to represent the only prompt word of this category.

WARP [8]: The model uses continuous vectors instead of discrete words to represent the categories. The output of the [MASK] position also obtains its hidden vector, and the two calculate the probability of belonging to different categories through the dot product. In the experiment, we use the word embedding of the category word as the initialization of the category vector.

PETAL [22]: The model uses labeled data and unlabeled data to automatically search for prompt words from PLM pruned vocabularies. By maximizing the likelihood function, it ultimately prefers to select words with higher frequency.

Auto-L [17]: The model sequentially prunes the search space through the initial probability distribution of the vocabulary and maximizing the accuracy in the zero-shot task and finally uses reordering to search for the best top n prompt words on the validation set. We fixed the automatic template generation part of the model and only use the construction part of the verbalizer.

KPT [23]: This method expands the verbalizer with the help of external knowledge and then refines the selected prompt words in various ways on the support set.

4.3 Experiment settings

The PLMs in the model adopt RoBERTa large. For zero-shot experiments, since there are no trainable parameters, we use the results of one experiment as the experimental data. For few-shot experiments, we use five different seeds to randomly select data, and the final experimental data are obtained by averaging the results from these five experiments. This setting ensures that the experimental findings are not overly influenced by a specific random initialization and provide a more robust and reliable assessment of the model's performance. Macro F1 is used as the test indicator in the experiment.

4.4 Main results

Table 3 contains all the experimental results on the four datasets, where AVG represents the average performance of each model of the

TABLE 3 Macro F1 (%) of different models on the four datasets.

K	Dataset	Finetuning	Manual	WARP	PETAL	Auto-L	KPT	Ours
0-shot	Restaurant	5.4	28.8	–	–	–	38.1	**74.4**
	Boots	15.9	32.4	–	–	–	28.8	**34.7**
	Keyboards	11.3	22.7	–	–	–	20.4	**25.1**
	TV	3.4	18.7	–	–	–	16.3	**26.2**
	AVG	9.0	25.7	–	–	–	25.9	**33.4**
5-shot	Restaurant	40.3	67.6	70.9	63.2	71.2	67.9	**73.5**
	Boots	23.7	55.4	60.1	48.6	57.2	55.6	**60.9**
	Keyboards	22.2	39.6	39.7	42.8	44.3	40.1	**45.4**
	TV	25.9	47.9	44.1	46.3	49.7	47.8	**51.2**
	AVG	28.0	52.6	53.7	50.2	55.6	52.9	**57.8**
10-shot	Restaurant	66.5	70.3	72.2	76.5	78.0	77.3	**78.8**
	Boots	43.2	61.6	60.2	48.4	58.3	66.1	**67.2**
	Keyboards	30.2	49.6	**51.4**	43.2	45.6	46.3	51.3
	TV	43.7	48.6	47.5	46.8	50.6	49.2	**52.6**
	AVG	45.9	57.5	57.8	53.7	58.1	59.7	**62.5**
20-shot	Restaurant	78.4	79.2	76.6	80.3	80.6	81.2	**82.8**
	Boots	55.7	68.3	65.3	64.4	64.3	65.2	**69.2**
	Keyboards	44.1	60.1	58.8	56.2	56.5	57.4	**60.9**
	TV	51.9	50.9	52.2	50.1	51.8	51.1	**53.5**
	AVG	57.5	64.6	63.2	62.8	63.3	63.7	**66.6**

Bold values indicate the best performance.

four datasets and bold represents the optimal performance. As shown in the table, our model achieves almost the best results under all settings. Compared with the second-best model, our model increased by 9.3%, 5.9%, 4.7%, and 9.9%, respectively, on the four datasets under the zero-shot setting, and the growth rate was particularly obvious. It shows that the prompt words searched from the PLM vocabulary using our method can better represent the category labels. Under different K values of the few-shot task, our method maintains a certain degree of performance growth in different field datasets which indicates that our model has a certain degree of generalization. Using the average performance AVG for comparison, our model increased by 2.2%, 2.8%, and 2.0%, respectively, under different K-shots compared to the second-best model. This shows that introducing weights for each prompt word and further training are beneficial to the optimization of the mapping process.

When further comparing different prompt learning methods, it can be found that our model almost achieved the best results under all K value settings, which proved the effectiveness of the design of this method. When the K value is small, the effect of the PETAL model is lower. According to the construction method of the verbalizer, it is speculated that PETAL needs training data to search for prompt words. So, when the labeled data are less, the deviation of the searched prompt words is greater. Auto-L may not consider the word confusion problem, so the effect is still lacking. As the K value continues to increase, KPT becomes the best model among all baselines, proving that the model requires training data to reduce the impact of noise words to a certain extent.

In addition, it is observed that the finetuning method is lower than all cue learning models in both zero-shot and few-shot tasks, so prompt learning is an advantageous method when there is less labeled data. However, as the training data increases, that is, as the K value increases, the gap between the two results decreases. It can be speculated that when the K value increases to a certain value, the finetuning method will still show comparable results.

4.5 Ablation study

To evaluate the impact of some designs in the model on the final performance, we conduct ablation experiments. We tested the influence of the three parts of the indicating filtering mechanism on the four datasets, respectively, and the results are shown in Figure 2. “w/o pos,” “w/o spe,” and “w/o sem” mean not to use B^{pos}, B^{spe}, and B^{sem}, respectively for verbalizer filtering.

Compared with the complete model, the significant decrease in experimental results of three ablation models illustrates that these three parts of the indicator mechanism can greatly ensure

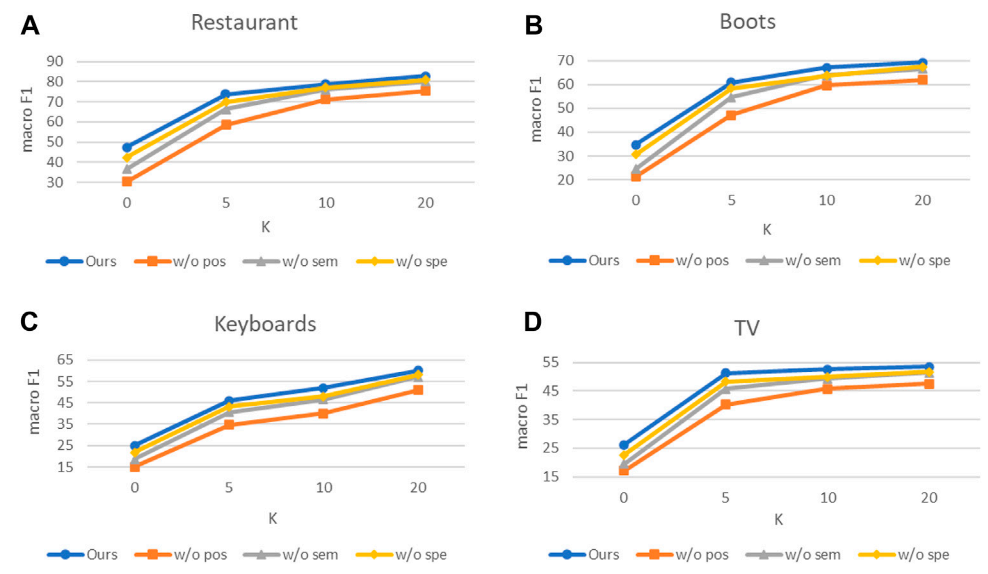


FIGURE 2
Ablation study on four datasets. **(A)** Ablation Study on "Restaurant" dataset. **(B)** Ablation Study on "Boots" dataset. **(C)** Ablation Study on "Keyboards" dataset. **(D)** Ablation Study on "TV" dataset.

that the most reasonable prompt words are searched for each category, thereby ensuring model performance. In addition, the following can be clearly observed: 1) The "w/o pos" model performs the worst on all four datasets, and the growth rate is lower than that of other models. This shows the prompt word set that has not been denoised contains more meaningless tokens, and these tokens have a higher prediction probability when filling in the [MASK] position, resulting in a decrease in the mapping performance of the verbalizer. 2) The performance of the "w/o sem" and "w/o spe" models is similar, indicating that category specificity and category semantic similarity are equally important when searching for prompt words. The common constraints of the two make each prompt word set not only have as many prompt words as possible and avoid mapping contradictions between different categories, which is beneficial to the subsequent mapping process.

4.6 Comparison of the miscellaneous category

This section quantitatively and qualitatively studies the effects of different models on the "miscellaneous" and "general" categories. The Amazon dataset contains the "general" category. For convenience of presentation, the two labels are collectively referred to as "miscellaneous" below. Figure 3 shows the results of each model under zero-shot and few-shot conditions, respectively. Table 4 shows the prompt words of "miscellaneous" obtained by different models. As shown in Figure 3, our model showed excellent results in different settings; especially in the zero-shot task, the improvement effect is obvious. On the zero-shot task, our method demonstrates improvements of 14.1%, 10.6%, 6.9%, and 11.1% compared to the second-best model across four datasets. Additionally, for the 10-shot task, our method exhibits enhancements of 3.0%, 4.8%, 6.3%, and 3.8% on the same datasets, respectively.

Referring to the data in Table 4, we speculate that because the sentences of the "miscellaneous" category have no obvious characteristics and the range of semantic expression is wide, the manual method only uses category word as the prompt word, which obviously cannot cover all data of this category, so the results are not ideal. Although KPT has expanded the scope of mapping, most of the prompt words searched from the external knowledge base for this category are uncommon and cannot be recognized by PLMs, resulting in poor performance in this category. Although the search-based model does not suffer from these two problems, it ignores the confusion between categories and can easily cause misjudgments during the prediction process. In addition, our model focuses on and solves the above problems, and the obtained prompt words have a high correlation with the category and can show good prediction ability on semantically ambiguous sentences.

4.7 Impact of the semantic extension

Our method still has certain prediction ability in the case of zero-shot because of using the semantic extended sentences of category labels as prior knowledge. This section studies the impact of the semantic extended sentences of category labels. Figure 4 shows the effect of the length of the semantic extension sentence on the final results. Wikipedia has a very detailed explanation for category words, usually from different aspects, so the optional range of semantic expansion sentences is long. The length "len" is calculated based on the number of tokens. In addition to using the category word itself with "len" as 1, the length of the semantic extended sentence is changed by continuously increasing the number of tokens in the definition statement.

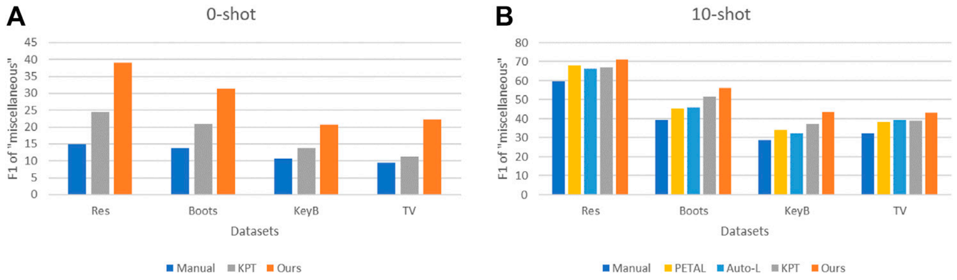


FIGURE 3
Experiments on the "Miscellaneous" category. **(A)** Zero-shot experiments on "Miscellaneous" category. **(B)** Few-shot experiments on "Miscellaneous" category.

TABLE 4 Prompt words for the "Miscellaneous" category.

Dataset	Method	Verbalizer token
Restaurant	Manual	miscellaneous
	PETAL	darkness, fiction, opinions, academia, interests, sociology, links,. . .
	Auto-L	Bryan, anonymous, Wes, noise, LM, Ethan, Wayne, dark,. . .
	KPT	heterogeneous, diverse, dissimilar, disparate, different, unlike,. . .
	Ours	same, general, main, particular, whole, specific, various, primary,. . .
Amazon Boots	Manual	general
	PETAL	remembered, Articles, arrived, finished, published, instructed,. . .
	Auto-L	produced, systems, published, female, default, quoted, . . .
	KPT	army, officer, brigadier, military, air, commander, field,. . .
	Ours	interesting, done, closed, clear, true, possible, established, like,. . .
Amazon Keyboards	Manual	general
	PETAL	votes, remarks, guy, Subject, excerpt, speakers, policy,. . .
	Auto-L	god, voice, journal, Jackson, guy, James, blogger, admin, hi,. . .
	KPT	lieutenant, cosmopolitan, universal, ecumenical, consumable,. . .
	Ours	included, fix, changed, summary, various, basic, likely,. . .
Amazon TV	Manual	general
	PETAL	url, AUTHOR, Hannah, username, starred, published, Votes,. . .
	Auto-L	controversy, followers, community, ranking, Society, twitter,. . .
	KPT	generality, rank, oecumenical, commander, admiral, full. . .
	Ours	titled, defined, concluded, called, summarized, cited, listed,. . .

Combining the results of the four datasets, it can be observed that the experimental results have improved with the increase in extended sentence tokens. This may be because the semantics of sentence expressions are rich, and PLMs can better understand the meaning of category labels to search for more reasonable prompt words. However, when the length is too long, the performance decreases instead. We speculated that it may be because the meaning contained in the semantic extensions is too complex leading to understanding deviation, which is not conducive to the model to choose more accurate prompt words. The semantic extended sentences used in the best experimental results of the method can be found in Table 2.

4.8 Impact of the templates

Template is another important component that affects prompt learning performance, so in this section, we tested the impact of different prompt templates on the proposed method. Table 5 lists all

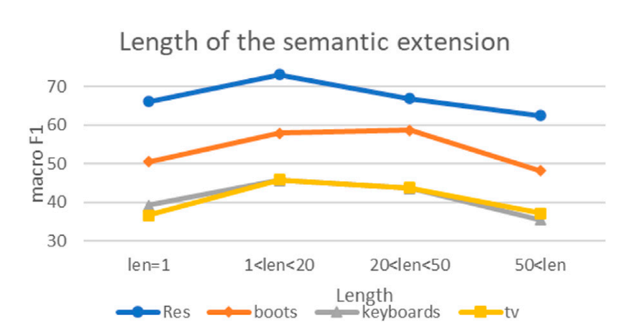


FIGURE 4
Impact of semantic extension length.

TABLE 5 Templates used in experiments.

ID	Templates
1	The "mask" category is discussed
2	The sentence discusses the "mask" category
3	It is about the "mask" category
4	Category: "mask"

the templates used in experiments. Figure 5 shows the results of using different templates on the Restaurant and Boots datasets under the 10-shot setting. As can be seen from the figure, our model not only maintains excellent performance in both datasets but also has a relatively gentle change curve compared with some other methods, indicating that it has a certain degree of robustness to different templates.

4.9 Impact of hyperparameters

In this section, we explore the impact of hyperparameters on experimental results and conduct grid searches on the Restaurant and Boots datasets for the two hyperparameters of "taking the first M words" and "taking the specificity probability greater than the threshold α." For the parameter M and parameter α, we set them to $\{50, 100, 300, 500, 1000\}$ and $\{0.90, 0.80, 0.75, 0.70, 0.60\}$, respectively. We use grid search to find the optimal values within the ranges of two parameters. The experimental results for parameter M are shown in Figure 6A. The results show that as the M value continues to increase, the model performance increases. However, when M increases to 1,000, the performance decreases, indicating that at too large M, it may select some low-quality prompt words, resulting in the reduction of the final classification results. Similar to the M value, as shown

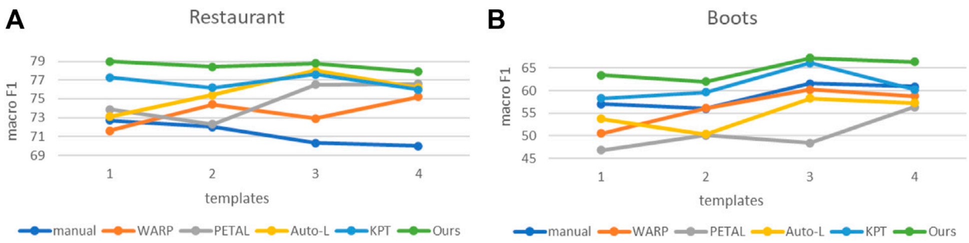


FIGURE 5
Impact of templates. **(A)** Experiment on "Restaurant" dataset. **(B)** Experiment on "Boots" dataset.

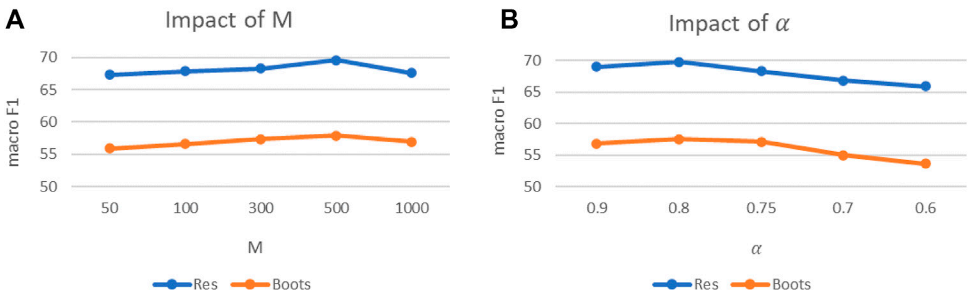


FIGURE 6
Impact of hyperparameters. **(A)** Experiment of parameter M. **(B)** Experiment of parameter α.

TABLE 6 Case study.

Examples	Manual prompt	KPT	Ours
I highly recommend this restaurant!! (Restaurant dataset)	food (✗)	food (✗)	miscellaneous (✓)
If you turn backlight all the way down it gets better (TV dataset)	sound (✗)	image (✓)	image (✓)

in Figure 6B, the α experimental curve also shows a trend of increasing first and then decreasing. This is because a too small α value will also introduce low-quality words and affect the model performance. The best experimental results in this article were obtained when M = 800 and α=0.8.

4.10 Case study

Table 6 shows some examples from different test sets. For example 1, the meaning expressed by this sentence does not belong to the categories "food," "service," "price," and "ambience," but to "miscellaneous." We speculate that due to the word "restaurant" in the sentence, the prompt word "restaurant" in the "food" category from the KPT model is easy to obtain a higher probability, and thus, it is mapped to the "food" category. The manual method detected errors in both examples. This may be because the category words themselves cannot better summarize the meaning of the example sentences, and it is easy to be misjudged as other categories.

5 Conclusion

In this paper, we propose a simple and effective method for aspect category detection based on prompt learning. To address the challenge of lack of labeled data and external knowledge, the semantic expansion of category labels is exploited to build the initial verbalizer. Additionally, we employ an indication mechanism to construct an appropriate verbalizer for category mapping. We conduct experiments on zero-shot and few-shot settings, respectively, and the results demonstrated the superiority of the proposed method. In our article, the verbalizer is constructed under a predefined manual template. In recent years, there has been a lot of work exploring the design of templates, but in most cases, the construction of the two is still separated, and both require certain labeled data. Therefore, in future work, we plan to further explore how to build prompt templates and verbalizers simultaneously to find the best combination of these two components.

.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

YX: conceptualization, formal analysis, methodology, validation, and writing–review and editing. YQ: software, validation, and writing–original draft. YX: conceptualization, formal analysis, methodology, validation, and writing–review and editing.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62162037 and U21B2027), Yunnan provincial major science and technology special plan projects (Grant Nos. 202302AD080003 and 202303AP140008), and Yunnan provincial major science and technology special plan projects (Grant No. 202301AT070444).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Hu M, Zhao S, Guo H, Xue C, Gao H, Gao T, et al. *Multi-label few-shot learning for aspect category detection* (2021). *arXiv preprint arXiv:2105.14174*.

2. Zhou X, Wan X, Xiao J. Representation learning for aspect category detection in online reviews. In: Proceedings of the AAAI conference on artificial intelligence, 29 (2015). p. 1547–52. doi:10.1609/aaai.v29i1.9194

3. Xue W, Zhou W, Li T, Wang Q. Mtna: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (2017). p. 151–6. 2: *Short Papers*.

4. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. *Language models are few-shot learners advances in neural information processing systems* (2020). 33.

5. Petroni F, Rocktäschel T, Lewis P, Bakhtin A, Wu Y, Miller AH, et al. *Language models as knowledge bases?* (2019). *arXiv preprint arXiv:1909.01066*.

6. Schick T, Schütze H. *Exploiting cloze questions for few shot text classification and natural language inference* (2020). *arXiv preprint arXiv:2001.07676*.

7. Kiritchenko S, Zhu X, Cherry C, Mohammad S. Nrc-Canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th international workshop on semantic evaluation (2014). p. 437–42. SemEval 2014.

8. Hambardzumyan K, Khachatrian H, May J. *Warp: word-level adversarial reprogramming* (2021). *arXiv preprint arXiv:2101.00121*.

9. Xenos D, Theodorakakos P, Pavlopoulos J, Malakasiotis P, Androutsopoulos I. Aueb-absa at semeval-2016 task 5: ensembles of classifiers and embeddings for aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (2016). p. 312–7. *SemEval-2016*.

10. Hercig T, Brychcín T, Svoboda L, Konkol M. Uwb at semeval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (2016). p. 342–9. Sem Eval-2016.

11. Toh Z, Su J. Nlangp at semeval-2016 task 5: improving aspect based sentiment analysis using neural network features. In: Proceedings of the 10th international workshop on semantic evaluation (2016). p. 282–8. *SemEval-2016)*.

12. Su Q, Xiang K, Wang H, Sun B, Yu S. Using pointwise mutual information to identify implicit features in customer reviews. In: International Conference on Computer Processing of Oriental Languages. Springer (2006). p. 22–30.

13. Schouten K, Van Der Weijde O, Frasincar F, Dekker R. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE Trans cybernetics* (2017) 48:1263–75. doi:10.1109/tcyb.2017.2688801

14. Jiang Z, Xu FF, Araki J, Neubig G. How can we know what language models know? *Trans Assoc Comput Linguistics* (2020) 8:423–38. doi:10.1162/tacl_a_00324

15. Haviv A, Berant J, Globerson A. *Bertese: learning to speak to bert* (2021). *arXiv preprint arXiv:2103.05327*.

16. Shin T, Razeghi Y, Logan IV RL, Wallace E, Singh S. *Autoprompt: eliciting knowledge from language models with automatically generated prompts* (2020). *arXiv preprint arXiv:2010.15980*.

17. Gao T, Fisch A, Chen D. *Making pre-trained language models better few-shot learners* (2020). *arXiv preprint arXiv:2012.15723*.

18. Li XL, Liang P. *Prefix-tuning: optimizing continuous prompts for generation* (2021). *arXiv preprint arXiv:2101.00190*.

19. Lester B, Al-Rfou R, Constant N. *The power of scale for parameter-efficient prompt tuning* (2021). *arXiv preprint arXiv:2104.08691*.

20. Qin G, Eisner J. *Learning how to ask: querying lms with mixtures of soft prompts* (2021). *arXiv preprint arXiv:2104.06599*.

21. Han X, Zhao W, Ding N, Liu Z, Sun M. Ptr: prompt tuning with rules for text classification. *AI Open* (2022) 3:182–92. doi:10.1016/j.aiopen.2022.11.003

22. Schick T, Schmid H, Schütze H. *Automatically identifying words that can serve as labels for few-shot text classification* (2020). *arXiv preprint arXiv:2010.13641*.

23. Hu S, Ding N, Wang H, Liu Z, Wang J, Li J, et al. *Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification* (2021). *arXiv preprint arXiv:2108.02035*.

24. Gu Y, Han X, Liu Z, Huang M. *Ppt: pre-trained prompt tuning for few-shot learning* (2021). *arXiv preprint arXiv:2109.04332*.

25. Zhu Y, Wang Y, Qiang J, Wu X. Prompt-learning for short text classification. *IEEE Trans Knowledge Data Eng* (2023) 1–13. doi:10.1109/tkde.2023.3332787

26. Cui G, Hu S, Ding N, Huang L, Liu Z. *Prototypical verbalizer for prompt-based few-shot tuning* (2022). *arXiv preprint arXiv:2203.09770*.

27. Meng Y, Zhang Y, Huang J, Xiong C, Ji H, Zhang C, et al. *Text classification using label names only: a language model self-training approach* (2020). *arXiv preprint arXiv:2010.07245*.

28. Liu H, Zhang F, Zhang X, Zhao S, Sun J, Yu H, et al. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022). p. 1079–87.

29. Zhang W, Song X, Feng Z, Xu T, Wu X. *Labelprompt: effective prompt-based learning for relation classification* (2023). *arXiv preprint arXiv:2302.08068*.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 20 March 2024
DOI 10.3389/fphy.2024.1388364

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Yafei Zhang,
Kunming University of Science and Technology, China
Zhiyuan Zhu,
Southwest University, China
Xing Li,
Nanjing University of Aeronautics and Astronautics, China

*CORRESPONDENCE
Li Yin,
✉ yl1@cqu.edu.cn
Li Tang,
✉ tangli_whu@163.com

RECEIVED 19 February 2024
ACCEPTED 06 March 2024
PUBLISHED 20 March 2024

CITATION
Chen H, Li Z, Huang X, Peng Z, Deng Y, Tang L and Yin L (2024), SCSONet: spatial-channel synergistic optimization net for skin lesion segmentation.
Front. Phys. 12:1388364.
doi: 10.3389/fphy.2024.1388364

SCSONet: spatial-channel synergistic optimization net for skin lesion segmentation

Haoyu Chen[1], Zexin Li[1], Xinyue Huang[1], Zhengwei Peng[1], Yichen Deng[1], Li Tang[2,3]* and Li Yin[2,3]*

[1]College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]Department of Radiation Oncology, Chongqing University Cancer Hospital, Chongqing, China, [3]Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital, Chongqing, China

In the field of computer-assisted medical diagnosis, developing medical image segmentation models that are both accurate and capable of real-time operation under limited computational resources is crucial. Particularly for skin disease image segmentation, the construction of such lightweight models must balance computational cost and segmentation efficiency, especially in environments with limited computing power, memory, and storage. This study proposes a new lightweight network designed specifically for skin disease image segmentation, aimed at significantly reducing the number of parameters and floating-point operations while ensuring segmentation performance. The proposed ConvStem module, with full-dimensional attention, learns complementary attention weights across all four dimensions of the convolution kernel, effectively enhancing the recognition of irregularly shaped lesion areas, reducing the model's parameter count and computational burden, thus promoting model lightweighting and performance improvement. The SCF Block reduces feature redundancy through spatial and channel feature fusion, significantly lowering parameter count while improving segmentation results. This paper validates the effectiveness and robustness of the proposed SCSONet on two public skin lesion segmentation datasets, demonstrating its low computational resource requirements. https://github.com/Haoyu1Chen/SCSONet.

KEYWORDS

light-weight model, medical image segmentation, attention mechanism, mobile health, skin lesion segmentation

1 Introduction

In 2024, it is projected that around 99,700 cases of *in situ* melanoma will be diagnosed, with an estimated 13,120 deaths from skin cancer, of which melanoma accounts for 99% [1]. Early detection of melanoma can often lead to cure through simple outpatient surgery, as opposed to late-stage diagnosis significantly reducing survival rates from over 99%–32%. Early detection is thus crucial for improving survival chances [2].

Dermatologists typically use dermatoscopy, an intuitive method for skin lesion examination, which relies on experienced doctors manually inspecting images [3]. However, this method can be less accurate for inexperienced dermatologists [4].

Traditional image segmentation methods, such as threshold-based [5], edge-based [6], and clustering-based [7] approaches, have played a role but are often time-consuming and error-prone, with limited effectiveness on complex datasets. In contrast, deep learning enhances accuracy and adaptability in image segmentation, making skin disease diagnosis more efficient and widespread.

Over the years, with the enhancement of computing capabilities and advancements in deep learning technologies, segmentation methods based on convolutional neural networks have seen significant performance improvements [8]. Fully Convolutional Networks (FCN) were developed as pioneers for semantic segmentation [9]. The introduction of the U-Net network marked a major breakthrough in medical image segmentation [10]. Following that, the integration of Transformer technology through Vision Transformer (ViT) further enhanced the capabilities in medical image segmentation [11]. These advanced network technologies continue to push the performance and accuracy of medical image segmentation, providing more efficient and widespread technical support for the diagnosis of skin diseases.

Previous work on enhancing the performance of the U-Net network has tended to introduce more complex modules. However, in the field of medical image segmentation, the importance of model lightweighting is self-evident. In the modern medical field, especially in the application of medical image analysis, the importance of lightweight models is becoming increasingly prominent. These models can run efficiently on devices with limited memory and processing capabilities, and they show great potential in mobile healthcare and rapid response scenarios. Forn make high-qua instance, in emergencies, they can be used to quickly diagnose a patient's condition, saving valuable treatment time. Moreover, these models are particularly valuable in remote areas because they cality medical diagnostic services more widespread and accessible, representing a significant advancement for typically resource-poor regions.Furthermore, the economic benefits of lightweight models cannot be overlooked. They can reduce the investment in hardware and operations for medical institutions, bringing cost benefits to medical systems around the world, especially in developing countries. By lowering medical costs, lightweight models provide more equal opportunities for medical services to a broader population, thereby helping to address socio-economic inequalities. In summary, the development of lightweight medical image segmentation models is not only a manifestation of technological progress but also an important part of social responsibility and commitment, aiming to improve the health level of all humanity by popularizing high-quality medical services.

To address the need for lightweight models, solutions like MobileNets [12–14] and transformer-based lightweight models such as MobileViT [15] have been developed for real-time image classification and segmentation of 2D images. Inception-ResNet optimizes inception modules and residual networks to enhance image feature extraction and detail restoration [16]. Additionally, in medical image segmentation, MISegNet [17] offers a powerful yet lightweight network for real-time segmentation of multimodal medical images. The UNeXt [18] model, combining UNet and MLP technologies, reduces parameters and computational load while maintaining high performance. MALUNet, through channel reduction and multiple attention mechanisms, shows superior performance in skin lesion segmentation, maintaining compactness and efficiency [19].

While existing lightweight medical image segmentation models have made progress in reducing computational resource consumption, they often overlook the issues of spatial and channel redundancy. Previous research has shown that there is considerable redundancy in both the spatial and channel dimensions of deep neural network feature maps. This redundancy can lead to insufficient extraction of key edge features in lesion areas, affecting the model's performance and segmentation accuracy. Moreover, the presence of redundancy leads to wasteful use of computational resources. Therefore, addressing spatial and channel redundancy is crucial for enhancing the segmentation performance of lightweight medical image models.

In this study, we designed a U-shaped network architecture, the core of which is the Spatial-Channel Fusion Block (SCF Block). In addition, by incorporating ConvStem at the initial stage of feature extraction, we combined the stability of traditional convolution with the dynamic adaptability of Omni-dimensional Dynamic Convolution (ODConv) [20]. Additionally, our network introduces Channel Attention Bridge Block (CAB) and Spatial Attention Bridge Block (SAB) through skip connections, effectively achieving fusion of multi-level and multi-scale information. The core SCF Block, based on Spatial and Channel Reconstruction Convolution (SCConv) [21], significantly reduces feature redundancy through spatial-channel feature fusion technology, incorporating the Efficient Multi-Scale Attention Module (EMA) [22] and Partial Convolution (Pconv) [23] to establish short and long-range dependencies and enhance feature extraction capabilities. This ensures a substantial improvement in SCConv's segmentation performance while reducing the parameter count and computational cost.In summary, our contributions are threefold:

- The Spatial-Channel Fusion Block (SCF Block) introduced aims to apply SCConv in the medical image segmentation field, reducing redundancy in the spatial and channel dimensions of feature mappings as well as in dense model parameters. It enhances the model's ability to extract key edge features in lesion areas, significantly reducing parameter count and computational cost while ensuring segmentation accuracy.
- We introduced a unique lightweight feature extractor, ConvStem, that employs the ODConv convolution mechanism. By learning four different types of attention in parallel across the four core spatial dimensions, this mechanism not only enhances the model's efficiency in capturing features but also significantly reduces the additional number of parameters. ConvStem merges the stability of traditional convolution with the dynamic adaptability of enhanced convolution structures, ensuring the model remains lightweight while effectively capturing a richer array of local features and details.
- We present SCSONet, a model characterized by innovative lightweight design and efficient feature extraction mechanisms. It significantly reduces the model's parameters while maintaining segmentation accuracy. This approach not only streamlines the computational demands but also enhances the model's applicability in real-world scenarios where resources are limited, ensuring both high performance and efficiency in medical image segmentation tasks.

2 Related work

In the evolution of medical image segmentation, Convolutional Neural Networks (CNN) have played a pivotal role. The introduction of Fully Convolutional Neural Networks (FCN) laid the foundation for precise segmentation and identification of target areas in images. UNet, with its encoder-decoder structure and efficient skip connections, significantly advanced medical image segmentation. Following UNet, architectures like 3D U-Net [24], V-Net [25], and U-Net++ [26] improved segmentation performance through enhanced convolution processes and connections. SF-Net [27] is an innovative multi-task framework that boosts tumor segmentation precision by fusing multimodal features and employing an uncertainty-based method for adaptive loss weight adjustment.TDGraphDTA Through multi-scale information interaction and graph optimization techniquesthe method enhances the accuracy of predictions and the interpretability of the model [28].BTSFDS-EI-MMRI [29] develops an advanced technique utilizing the Swin Transformer and CNNs for enhanced MRI image analysis, focusing on integrating semantic and edge features for improved accuracy.X-Net [30] combines CNNs and Transformers for improved medical image segmentation, employing a dual architecture for enhanced feature extraction and accuracy on small datasets.GSOMMIF-AL [31] introduces an adversarial approach for enhancing glioma segmentation from multi-modal MR images, emphasizing image fusion for better segmentation outcomes.MISMFIF [32] presents a cloud-enhanced medical image segmentation technique, integrating Transformers and CNNs for robust feature extraction and employing an interactive module for improved accuracy, showcasing cloud computing's scalability and performance advantages.ASTCMSeg [33] presents a 3-D self-training framework for segmenting medical images across different modalities without paired data, focusing on anatomical consistency and a novel frequency domain approach for improved accuracy.ViT-UperNet [34], a hybrid model leveraging vision transformers and a unified-perceptual-parsing network, excels in medical image segmentation by combining self-attention with multi-scale feature fusion, significantly improving accuracy on cardiac MRI images.

As models grow in scale and complexity, so do their computational and storage costs, limiting their practical application in resource-constrained settings. This highlights the need for optimizing model efficiency without compromising performance, ensuring that advanced medical image segmentation technologies can be effectively deployed in diverse environments, particularly where computational resources are scarce.

To address these challenges, researchers are focusing on the design of lightweight segmentation networks for efficient visual processing. Innovations such as MobileNets with depthwise (DW) and pointwise (PW) convolutions, grouped convolutions from AlexNet [35], ODConv with multidimensional attention, PConv focusing on reducing redundant computation, and SCConv reducing feature map redundancy, are paving the way for more resource-efficient and practical models in medical image segmentation, especially in scenarios with limited computational and memory resources.

Recently, UNeXt, based on Multi-Layer Perceptrons (MLP) and UNet, has become a more suitable solution for practical applications in medical image segmentation due to its significant reduction in parameter count. MALUNet, as a lightweight medical image segmentation model incorporating various attention mechanisms, is better suited for clinical settings. However, despite these advancements, lightweight models still have performance gaps compared to larger models, with room for improvement in parameter efficiency and GFlops. Additionally, these methods have not fully addressed the redundancy in spatial and channel dimensions during the feature extraction process.

Given these considerations, this paper introduces an innovative lightweight UNet segmentation model based on the Spatial-Channel Fusion Block (SCF Blcok). This model effectively addresses spatial and channel redundancy issues by fusing multi-level, multi-scale information in skip connection paths, simultaneously enhancing segmentation accuracy and efficiency. Its innovation lies in its ability to deliver efficient, accurate segmentation results while maintaining low computational complexity, making it an ideal choice for practical applications and the mobile health domain. This approach offers a more efficient, practical solution for medical image analysis, also providing new directions for future developments in medical image segmentation technology.

3 Methods

3.1 Overview

The proposed skin lesion segmentation framework is shown in Figure 1, which consists of ConvStem, SCF Block, and SCAB modules. ConvStem enhances flexibility and reduces parameters through dynamic adjustment of convolution kernels via omni-dimensional attention, moving beyond static application. The SCF Block, comprising SCConv for spatial-channel reconstruction, PConv, and EMA for establishing dependencies and enhancing feature extraction, reduces spatial and channel redundancy, refining feature representation. SCAB, with CAB and SAB, improves multi-level and multi-scale information fusion, reducing feature loss during downsampling.

3.2 ConvStem

Lesion areas in medical images often present irregular shapes and vary greatly across different images, making accurate identification and segmentation a highly challenging task. To enhance the performance of medical image segmentation, particularly in capturing fine-grained and shape-sensitive local details, this study introduces the ConvStem module, as shown in the top right corner of Figure 1.

Traditional Convolutional Neural Networks are limited in simulating complex and irregular shape changes due to the fixed geometric structure of their basic modules. To overcome this limitation, the ConvStem module employs ODConv at the initial feature extraction stage, an innovative convolutional method with Omni-Dimensional attention mechanisms. ODConv combines the stability of traditional convolution with the flexibility of dynamic

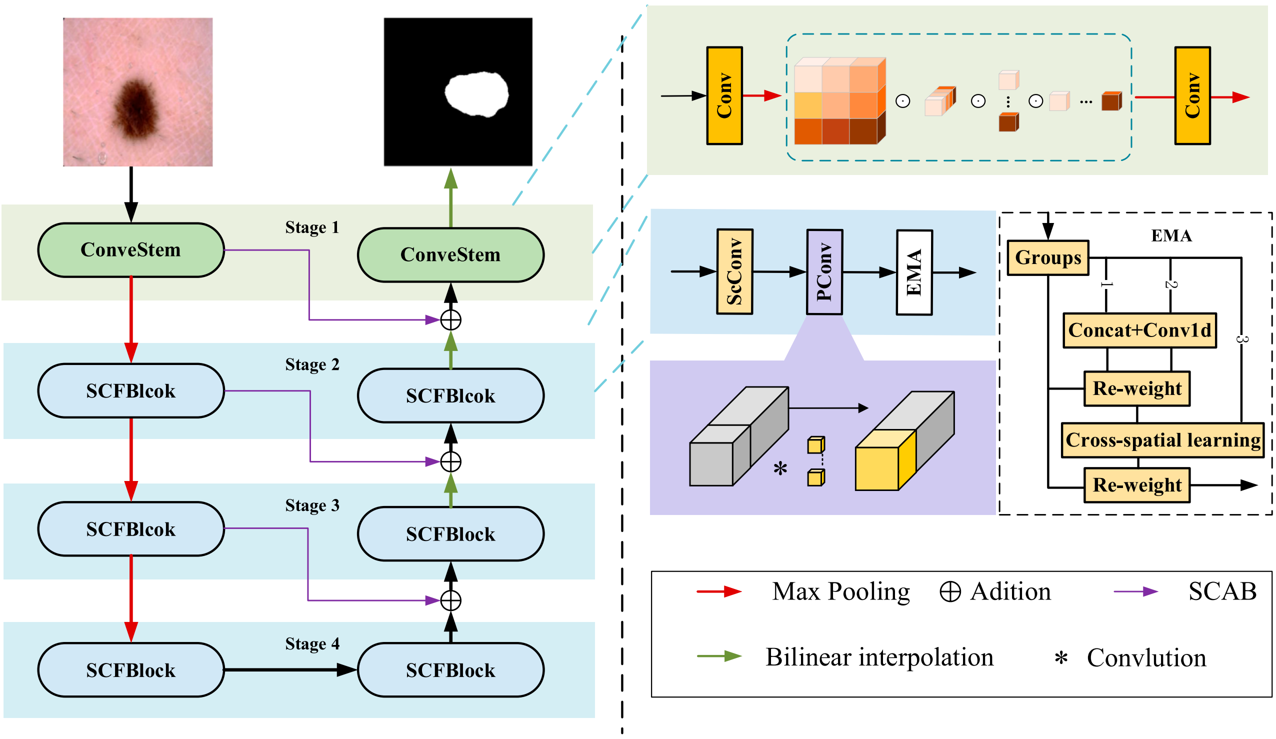


FIGURE 1
The framework of the proposed skin lsion segmentation method. It llustrates the framework for segmenting skin diseases, primarily consisting of three modules: the ConvStem module, the SCF Block, and the SCAB module for skip connections.

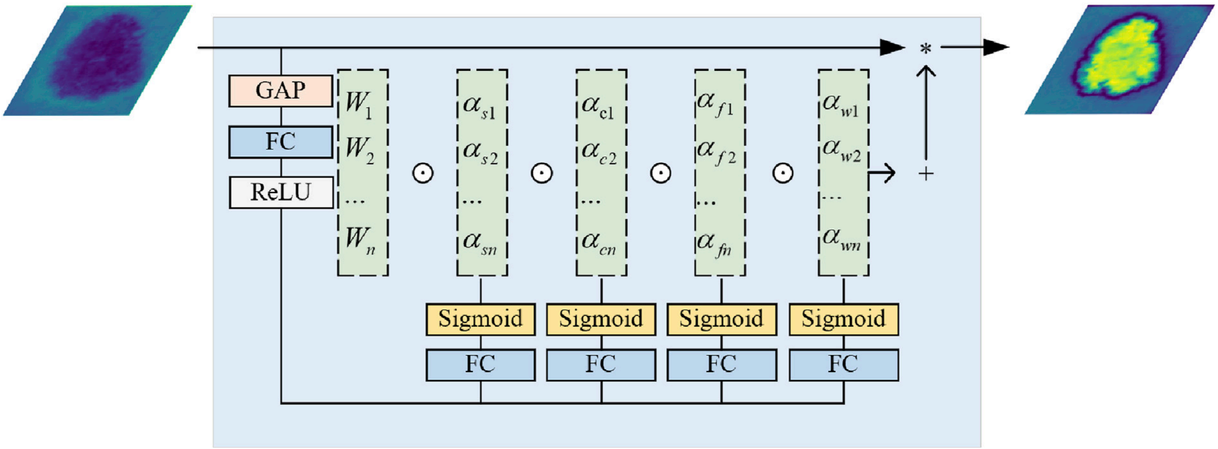


FIGURE 2
The architecture of Omni-dimensional Dynamic Convolution.

convolution, enabling the model to more effectively capture and process shape-aware features of irregularly shaped lesion areas in medical images. Through this design, the ConvStem module significantly enhances the network's ability to recognize complex lesion morphologies in medical images, thereby providing richer and more accurate primary feature information for subsequent network layers' feature learning and fusion.

The input image, denoted as $F_0 \in \mathbb{R}^{H\times W\times C}$, is initially operated by ConvStem.ConvStem consists of two standard convolutions and one ODConv, with a max pooling downsampling step in between. ODConv introduces a multidimensional attention mechanism that employs a parallel strategy to learn different attentions across all four spatial dimensions of the convolution kernel. Figure 2 provides a schematic illustration of ODConv, which can also be represented by the Eq. 1:

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right)^{*} x \tag{1}$$

where $\alpha_{wi} \in \mathbb{R}$ denotes the attention scalar for the convolutional kernel$W_i, \alpha_{\mathrm{si}} \in \mathbb{R}^{\mathrm{k}\times\mathrm{k}}, \alpha_{ci} \in \mathbb{R}^{C_{in}}$ and $\alpha_{fi} \in \mathbb{R}^{C_{out}}$ denote three newly introduced attentions, which are computed along the spatial dimension, the input channel dimension and the output channel dimension of the kernel space for the convolutional kernel W_i, respectively; denotes the multiplication operations along different dimensions of the kernel space. Here, α_{si}, α_{ci}, α_{fi} and α_{wi} are computed with a multi-head attention module.: (1)α_{si} assigns different attention scalars to convolutional parameters (per filter) at $k \times k$ spatial locations; (2)α_{ci} assigns different attention scalars to cin channels of each convolutional filter W_{mi}; (3) α_{fi} assigns different attention scalars to cout convolutional filters; (4) α_{wi}assigns an attention scalar to the whole convolutional kernel. ODConv enhances feature extraction focus and efficiency by dynamically concentrating on key aspects of the input features through its attention mechanism across each dimension.

This application within ConvStem allows the convolution kernels to adjust dynamically to different inputs, moving away from a static, singular approach. This increases the model's flexibility, reduces the number of parameters and computational burden, aiding in model lightweighting while boosting performance. Standard convolution captures basic features efficiently, while ODConv's dynamic adjustment provides a deeper understanding and extraction for specific features, crucial for complex skin disease image analysis. Combining these convolutions, ConvStem outputs feature mappings that finely reflect shapes and local details, enabling our proposed SCSONet to produce more detailed lesion segmentation results, showcasing rich, multi-faceted feature information. After passing through ConvStem, the input F_0 produces the outputs F_1, F_2 and F_3, as described by the following Eq. 2.

$$\begin{cases} F_1 = \mathrm{MaxPool}(\mathrm{Conv}(F_0)) \\ F_2 = \mathrm{MaxPool}(\mathrm{ODonv}(F_1)) \\ F_3 = \mathrm{MaxPool}(\mathrm{Conv}(F_2)) \end{cases} \tag{2}$$

F_1, F_2 and F_3 are each connected to the decoder through the SCAB, which includes a Channel Attention Bridge Block (CAB) and a Spatial Attention Bridge Block (SAB).

The SAB uses max and average pooling operations at each stage to establish short and long-range dependencies and enhance feature extraction capability. After these operations, feature maps with channel C, height H, and width W are concatenated into feature maps with two channels, while height and width remain unchanged. Dilated convolution and the sigmoid function are then applied to obtain spatial attention maps for each stage. Finally, these are element-wise multiplied with the initial images of the stage, and the residuals are summed, restoring the original channel count for each stage.

The CAB is primarily designed to fuse features across different channel orders to better integrate information. The internal workings of this module can be represented by the following Eq. 3:

$$\begin{cases} t_i' = GAP(t_i), \\ T = \mathrm{Concat}'(t_1', t_2', \ldots, t_{s-1}'), \\ T' = \mathrm{Conv}1D(T), \\ Att_i = \sigma(FC_i(T_i)), \\ \mathrm{Out}_i = t_i + t_i \odot Att_i. \end{cases} \tag{3}$$

P_i is the feature map obtained at stage input. Is the total number of stages, FC_i is the fully connected layer at stage, and σ is the sigmoid function.

The two bridge attention modules can fuse the multi-stage and multi-scale features of Stages 1–3 (Including the output from the subsequent SCF Block) to generate the attention maps in the spatial and channel dimension. And then, we add features obtained by bridge attention modules with features of the decoder part to reduce the feature semantic difference between the encoder and decoder while alleviating the information loss caused by the sampling process.

3.3 Spatial-channel fusion block (SCF block)

Although existing lightweight medical image segmentation models have made progress in reducing computational resource consumption, they often overlook issues of spatial and channel redundancy. To address these problems, the SCF Block was designed to significantly optimize the feature fusion process, particularly in reducing feature map redundancy across spatial and channel dimensions. By integrating the spatial-channel feature fusion technique of SCConv, the SCF module innovatively reduces feature redundancy, while the introduction of EMA and PConv enhances the model's ability to capture short and long-range dependencies, further improving the efficiency and accuracy of feature extraction. This method, which focuses on feature fusion, not only reduces computational costs and model parameters but also greatly enhances the quality and precision of the segmentation results while maintaining the model's lightweight stature. The SCF Block for Stage 2 can be represented as follows Eq. 4:

$$F_4 = \mathrm{MaxPool}(\mathrm{EMA}(\mathrm{PConv}(\mathrm{SCConv}(F_3)))) \tag{4}$$

As shown in Figure 3. SCConv initially obtains spatially refined features X_w through SRU operations, and then acquires channel-refined features Y using CRU operations.

The Spatial Reconstruction Unit (SRU) reconstructs redundant features based on weights to suppress redundancy in the spatial dimension and enhance feature representation. The formula for calculating weights is as follows Eq. 5:

$$W = \mathrm{Gate}\Big(\mathrm{Sigmoid}\Big(W_\gamma(GN(X))\Big)\Big) \tag{5}$$

The formula for reconstruct is as follows Eq. 6:

$$\begin{cases} X_1^w = W_1 \otimes X, \; X_2^w = W_2 \otimes X, \\ X_{11}^w \oplus X_{22}^w = X^{w1}, \; X_{21}^w \oplus X_{12}^w = X^{w2} \\ X^{w1} \cup X^{w2} = X^w. \end{cases} \tag{6}$$

The Channel Reconstruction Unit (CRU) employs a *Split* − *Transform* − *and* − *Fuse* strategy to reduce redundancy in the channel dimension, as well as computational and storage costs. After splitting, the spatially optimized feature X_w is divided into upper X_{up} and lower X_{low} parts. In the Transform stage, X_{up} undergoes efficient convolution operations (i.e., GWC and PWC), and the outputs are aggregated to form a combined representative feature map Y_1. The upper layer transformation stage can be represented as follows Eq. 7:

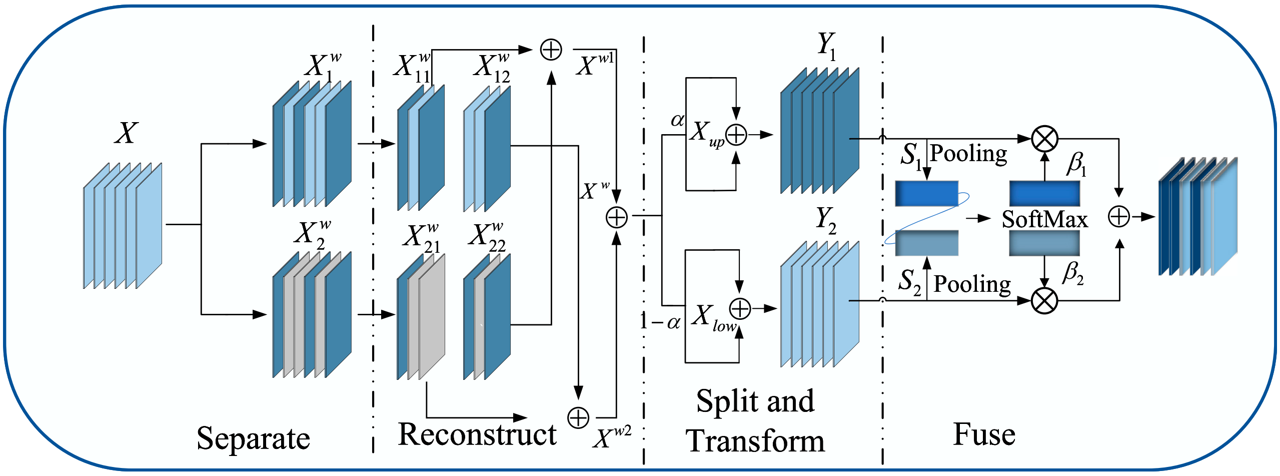


FIGURE 3
The architecture of Spatial and Channel Reconstruction Convolution.

$$Y_1 = M^G X_{up} + M^{P_1} X_{up} \tag{7}$$

$X_l ow$ is input into a lower transformation stage where a cost-effective 1 × 1 PWC operation is applied to generate a feature map Y_2 with shallow hidden details, complementing the rich feature extractor.

Global spatial information s_1 and s_2 are then collected through global average pooling, and channel soft attention operations produce feature importance vectors β_1 and β_2. These vectors guide the fusion of upper-layer features Y_1 and lower-layer features Y_2, generating refined features Y. The specific formula is as follows Eq. 8:

$$\begin{cases} \beta_1 = \dfrac{e^{s_1}}{e^{s_1} + e^{s_2}},\ \beta_2 = \dfrac{e^{s_1}}{e^{s_1} + e^{s_2}},\ \beta_1 + \beta_2 = 1 \\ Y = \beta_1 Y_1 + \beta_2 Y_2 \end{cases} \tag{8}$$

After passing through Partial Convolution (PConv), conventional convolution is applied to only a portion of the input channels for spatial feature extraction, with the remaining channels left unchanged, as shown on the right side of Figure 1. For continuous or regular memory access, the first or last continuous c_p channels are computed as a representation of the entire feature map. Therefore, the Floating Point Operations (FLOPs) of a PConv are significantly reduced, as indicated by the Eq. 9:

$$h \times w \times k^2 \times c_p^2 \tag{9}$$

With a typical partial ratio $r = \frac{c_p}{c} = \frac{1}{4}$, the FLOPs of a PConv is only $\frac{1}{16}$ of a regular Conv.To further reduce computational redundancy and ensure model lightweighting, on one hand, PConv applies convolution to only a subset of channels, refining channel features. On the other hand, the CRU in the subsequent SCF Block integrates the uneven channels from Partial Conv, achieving better performance with fewer parameters and more efficient computation, thus enhancing the overall quality of the features.

To address the potential loss of important features due to SCConv's spatial information compression, we introduced an Efficient Multi-Scale Attention (EMA) mechanism.

As shown on the right side of Figure 1, EMA reshapes part of the channel dimensions into batch dimensions, avoiding dimensionality reduction through standard convolution. This approach allows for different strategies in parallel subnetworks to maximally preserve multi-scale features of pathological sections. Moreover, EMA fuses output feature maps of two parallel subnets using a cross-space learning method, ensuring areas with potential targets in the final output feature map have higher feature weights.

By adjusting channel dimensions and applying multi-scale attention, EMA compensates for SCConv's limitations in feature extraction, enhancing the capture of key features, optimizing feature representation, and improving the accuracy of medical image segmentation, effectively addressing SCConv's limitations in handling fine-grained features.

The SCF Block represents a significant advancement in medical image segmentation, offering a robust solution for reducing redundancy while enhancing feature representation through spatial-channel fusion. By integrating SCConv, EMA, and PConv, it addresses the critical need for efficient, high-performance segmentation in medical imaging. This module's innovative approach to capturing fine-grained details and dependencies not only improves segmentation accuracy but also ensures the model's lightweight nature, making it an ideal choice for applications where computational resources are limited.

3.4 Loss function

In this study, each image in the dataset is associated with a corresponding binary mask. Skin lesion segmentation is treated as a pixel-level binary classification task, distinguishing skin lesions from the background. The combination of Binary Cross-Entropy (BCE) loss and Dice similarity coefficient loss is used as the loss function to optimize network parameters, effectively addressing the challenges of skin lesion segmentation by balancing pixel-wise accuracy and overlap between the predicted and ground truth masks.

The loss function is BceDice loss, which can be expressed by the Eq. 10:

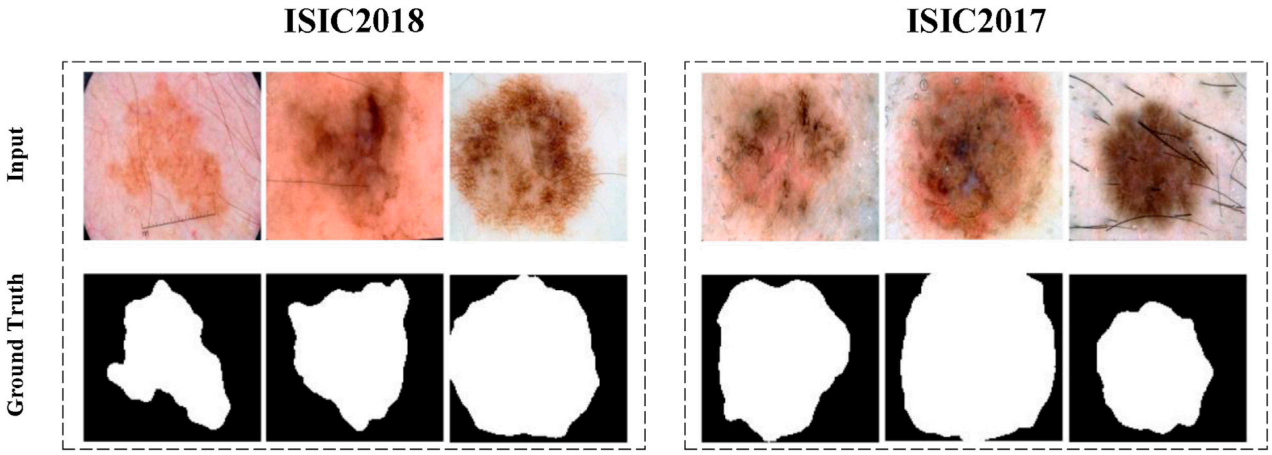


FIGURE 4
A portion of the ISIC2017 and ISIC2018 datasets.

$$\begin{cases} L_{Bce} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \\ L_{\text{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \\ L_{\text{BceDice}} = \alpha_1 L_{Bce} + \alpha_2 L_{\text{Dice}} \end{cases} \tag{10}$$

Where N is the total number of samples, y_i is the real label,p_i is the prediction. $|X|$ and $|Y|$ represent ground truth and $X \cap Y$ prediction, respectively. α_1 and α_2 refer to the weight of two loss functions. In this paper, both weights are taken as one by default.

4 Experiment

4.1 Datasets

The segmentation tasks were conducted on the ISIC2017 [36] and ISIC2018 datasets.Figure 4 showcasing a portion of the ISIC2017 and ISIC2018 datasets. The International Skin Imaging Collaboration (ISIC) dataset is a widely used open dataset in dermatological research. These datasets aim to facilitate computer-assisted dermatology diagnosis and research by providing a large collection of skin lesion images and related clinical metadata, supporting the development and validation of segmentation algorithms.

The ISIC2017 and ISIC2018 datasets contain 2,150 and 2,694 dermoscopic images with segmentation mask labels, respectively. For experimental purposes, the datasets were randomly split into training and testing sets at a 7:3 ratio. Specifically, the ISIC2017 dataset was divided into 1,500 images for training and 650 for testing, while the ISIC2018 dataset was divided into 1,886 images for training and 808 for testing. Comparative experiments were conducted on both ISIC2017 and ISIC2018, with ablation studies performed on ISIC2018.

4.2 Implementation details

All experiments were implemented in the PyTorch framework and conducted on an NVIDIA GeForce RTX 3070 Ti Laptop GPU with 8 GB of memory. Based on experience, all images were normalized and resized to 256 × 256, with data augmentation techniques including vertical flip, horizontal flip, and random rotation applied. The loss function used was the BceDice loss, represented by Eq. 10. AdamW was used as the optimizer with an initial learning rate of 0.001, employing a cosine annealing scheduler for learning rate adjustment, with a maximum of 50 iterations, a minimum learning rate of 0.00001, training epochs set to 300, and a batch size of 8.

Five metrics including Mean Intersection over Union (**mIoU**) and Dice similarity score (**DSC**), Eq. 11 are used to measure segmentation performances. In addition, Params is utilized to indicate the number of parameters, and the unit is Million (M). The computational complexity is calculated regarding the number of floating point operators (**GFLOPs**). Note that the parameters and GFLOPs of models are measured with 256 × 256 input size.

$$\begin{cases} \text{mIoU} = \frac{TP}{TP + FP + FN} \\ \text{DSC} = \frac{2TP}{2TP + FP + FN} \end{cases} \tag{11}$$

Where TP, FP, FN, TN represent true positive, false positive, false negative, and true negative.

4.3 Comparsion with other methods

In comparative experiments, the proposed SCSONet demonstrated significant advantages over advanced models like EGEUNet [37], showcasing its lightweight nature with fewer parameters and GFLOPs. Notably, SCSONet achieved the lowest GFLOPs among skin disease segmentation methods, at only 0.056, highlighting its efficiency. Figure 5 emphasized SCSONet's reduced computational demand, making it an ideal choice for resource-constrained environments while maintaining high segmentation performance.

Table 1 showcase SCSONet's performance against other methods on the ISIC2017 and ISIC2018 datasets, illustrating its

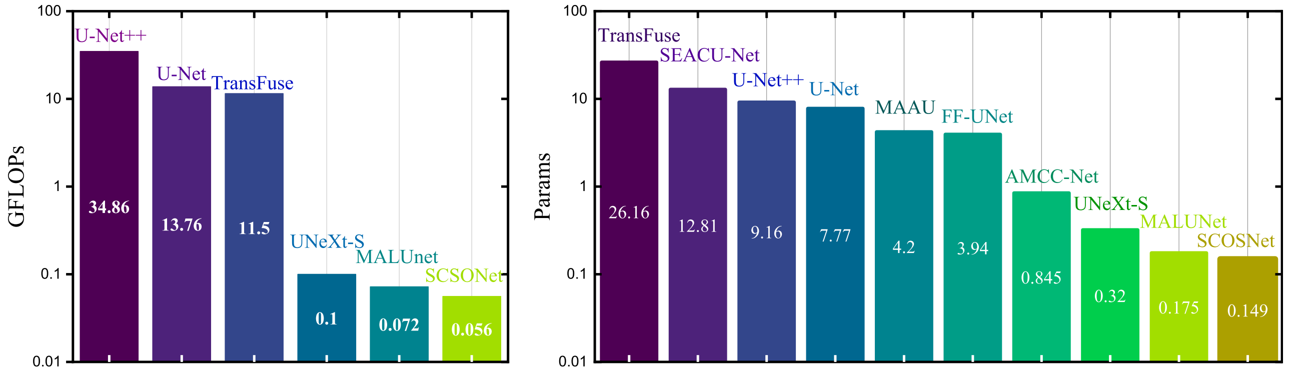


FIGURE 5
Histogram visualization with the *Y*-axis set as a logarithmic scale comparison with other methods on parameters and FLOPs.

TABLE 1 Comparative experimental results on the ISIC2017 and ISIC2018 dataset.

Data	Model	Params	GFLOPs	mIoU (%)	DSC(%)
ISIC2018	UNet (2015)	7.77	13.76	78.13	86.99
	Unet++(2018)	9.16	34.86	78.92	87.83
	TransFuse[38](2021)	26.16	11.5	80.63	89.27
	FF-UNet (2022)	3.94	—	80.2	88.7
	UNeXt-S (2022)	0.32	0.1	79.09	88.33
	MALUNet (2022)	0.175	0.083	80.25	89.04
	MAAU (2023)[39]	4.2	—	—	88.1
	AMCC-Net (2023)[40]	0.845	—	—	89
	SEACU-Net (2023)[41]	12.81	—	—	87.58
	EGE-Unet (2023)	**0.053**	0.072	80.94	89.46
	SCSONet (ours)	0.149	**0.056**	**80.99**	**89.5**
ISIC2017	UNet (2015)	7.77	13.76	76.98	86.99
	TransFuse (2021)	26.16	11.5	79.21	88.4
	UNeXt-S (2022)	0.32	0.1	78.26	87.8
	FAT-Net (2022)[42]	30	23	76.53	85
	MALUNet (2022)	0.175	0.083	78.78	88.13
	EGFNet (2022)[43]	0.52	—	—	84.87
	MMS-Net (2023)[44]	67.34	68.52	77.9	87.6
	QGD-Net (2023)[45]	0.777	—	72.58	—
	LCAUnet (2023)[46]	13.38	18.91	76.1	86.6
	EGE-Unet (2023)	**0.053**	0.072	79.81	88.77
	SCSONet (ours)	0.149	**0.056**	**80.14**	**88.97**

The bold values represent the optimal metrics.

state-of-the-art overall performance. Specifically, compared to larger U-net models, SCSONet not only achieved superior performance but also significantly reduced parameters and GFLOPs by 451× and 1,224×, respectively. It outperformed other lightweight models by increasing mIoU by 7.56% over QGD-Net, with fewer parameters. Surpassing EGEUNet, it demonstrated better results in mIoU and DCS while reducing GFLOPs by 22.2%, able to train within 0.6 GB of VRAM. Its effectiveness is showcased in Figures 5, 6.

The qualitative comparison results, as shown in Figure 7, involve randomly selected test samples for qualitative evaluation. It is observed

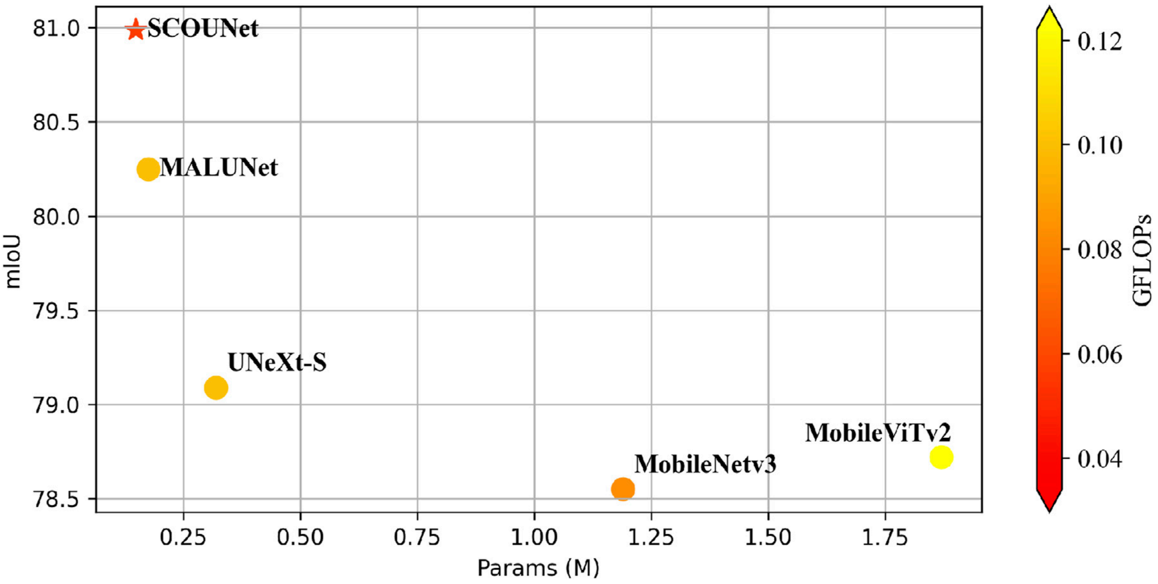


FIGURE 6
Lightweight model performance comparison.

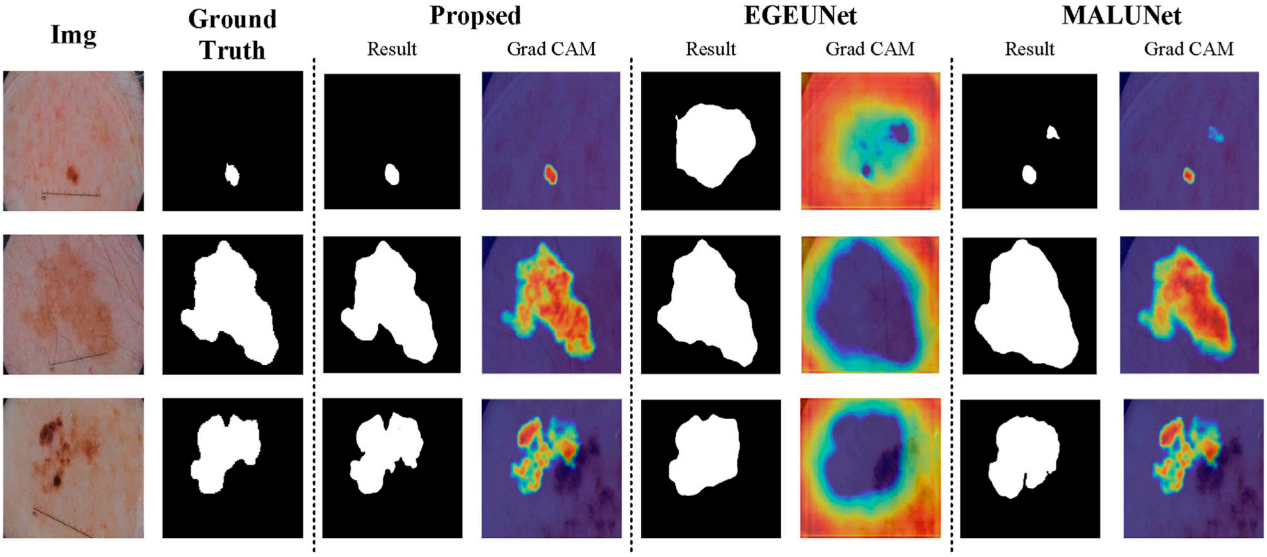


FIGURE 7
Comparison of segmentation results from different models on the ISIC2018 dataset and Grad CAM visualization (utilizing heatmaps to visualize the network prediction process.

that SCSONet effectively differentiates between skin lesion areas and normal skin, achieving more accurate target area localization and boundary prediction compared to other models, which show issues with over-segmentation and under-segmentation. These comparisons demonstrate SCSONet's effectiveness in skin lesion segmentation.

4.4 Ablation studies

As shown in Table 2, ablation studies were conducted to assess the effectiveness of each module within the proposed method. MALUNet served as the base model. Initially, ablation on the ConvStem module showed significant improvements in mIoU and DSC with notable reductions in parameters and GFLOPs, by replacing the first three convolutional layers in the base with ConvStem. Subsequently, replacing the base model's last three layers with three SCF Blocks similarly resulted in performance enhancement and reductions in parameters and GFLOPs. The ablation study meticulously demonstrates the significant contributions of key modules within SCSONet—ConvStem and SCF—towards enhancing medical image segmentation performance. The ConvStem module, by incorporating Omni-Dimensional Dynamic Convolution (ODConv), significantly

TABLE 2 Objective evaluation results of the ablation study on the ISIC2018 benchmark.

Model	Params	GFLOPs	mIoU (%)	DSC(%)
Base	0.175	0.083	79.01	88.27
Base + ConvStem	0.164	0.057	80.02	89.01
BASE + SCF Block	0.150	0.078	80.48	89.19
Base + ConvStem + SCF Block	0.149	**0.056**	**80.99**	**89.50**

The bold values represent the optimal metrics.

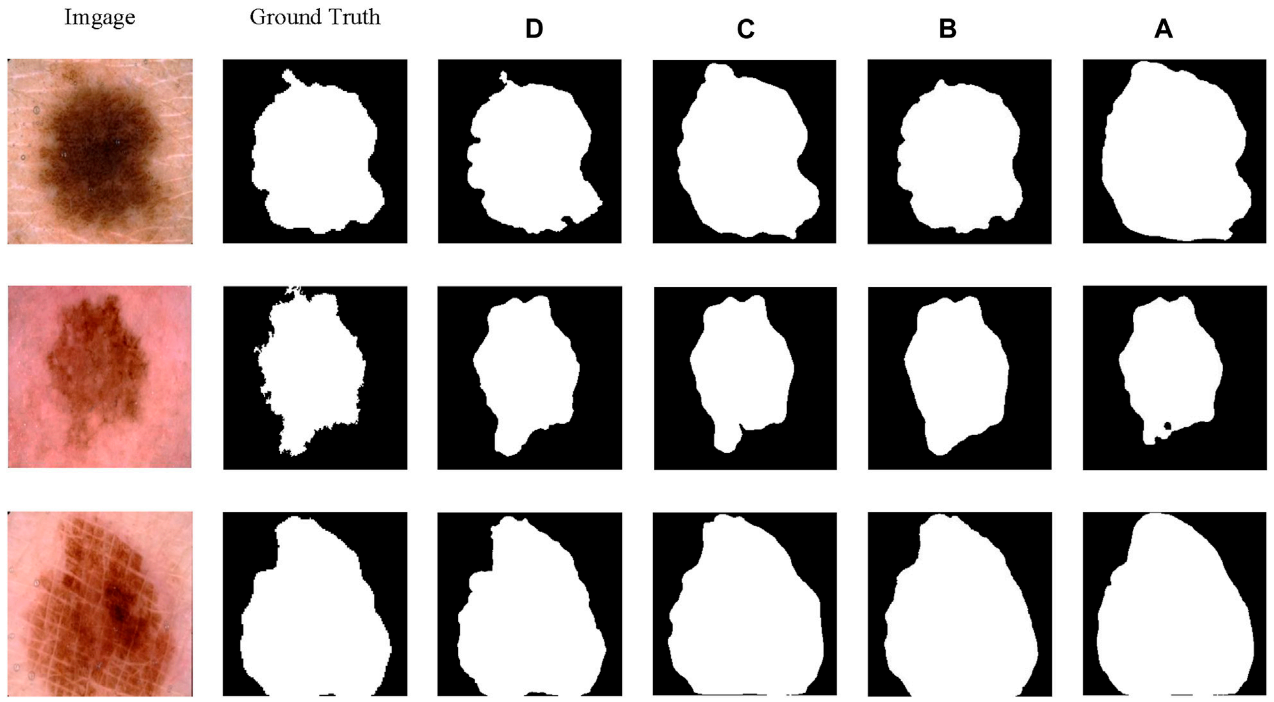


FIGURE 8
The results of the ablation study on the ISIC2018.

TABLE 3 Comparison and ablation experiments within the SCF Block.

Model	Params	GFLOPs	mIoU (%)	DSC(%)
SCConv	0.131	0.072	79.20	88.42
SCConv + PConv	0.148	0.075	80.24	89.01
SCConv + DepthwiseSeparableConv	0.141	0.074	79.86	88.75
SCConv + Dilated convolution	0.181	0.082	79.14	88.32
SCConv + PConv + EMA	0.149	**0.056**	**80.48**	**89.19**

The bold values represent the optimal metrics.

enhances the model's capability to recognize the shapes of irregular lesion areas, substantially improving the efficiency of primary feature extraction. Meanwhile, the SCF module effectively reduces feature map redundancy through spatial-channel feature fusion technology, further enhancing the model's segmentation precision and efficiency.Experimental results indicate that the inclusion of each module positively impacts model performance, particularly when used in combination, leading to optimal performance in terms of mIoU and DSC, while also achieving a significant reduction in the number of parameters and computational costs. These findings not only validate the effectiveness of the ConvStem and SCF modules in medical image segmentation tasks but also highlight the potential application of our lightweight network architecture in resource-constrained environments.Finally, for clearer visual comparison, experimental results are shown in Figure 8.

In Table 3, we conduct micro ablations on SCF Block.Further ablation studies within the SCF Block compared the effects of

PConv, DepthwiseSeparableConv, and Dilated convolution. The results highlighted PConv's significant contribution to SCConv's performance enhancement, also confirming the role of EMA within the SCF Block for improving the segmentation capabilities of the network.

SCSONet stands out as the first lightweight model to reduce GFlops to around 0.056 while maintaining exceptional segmentation performance. Its effectiveness is showcased in Figures 5, 6, which clearly present experimental results and segmentation outcomes, respectively. Demonstrating robust performance on two public datasets, SCSONet's primary clinical application is to assist in diagnosis, helping doctors quickly delineate focal areas or enabling non-specialists to diagnose diseases rapidly. Deploying this model in hospitals for semantic segmentation on small datasets can achieve higher segmentation accuracy.

5 Conclusion and future works

In the field of medical image processing and analysis, hospitals often rely on high-performance GPUs and large computational devices, requiring substantial computational resources. However, for under-resourced or remote medical facilities, limited computational resources pose a significant barrier to implementing advanced medical image analysis. This gap hinders the widespread adoption and application of advanced medical imaging technologies, especially in regions that need them most. And also, for rapid lesion detection and diagnosis in the field or emergency situations, a model that can be easily integrated into mobile devices is equally necessary.To address this challenge, this paper proposes SCSONet, an innovative lightweight network architecture comprising ConvStem, SCF Block, and skip connections, aimed at bridging this gap by enabling efficient, high-quality medical image analysis with lower computational demands.

The ConvStem module with full-dimensional attention effectively enhances the recognition of irregularly shaped lesion areas while reducing the model's parameter count and computational load, facilitating model lightweighting and performance improvement. The SCF Block, through spatial and channel feature fusion, efficiently reduces feature redundancy, significantly lowering parameter count while improving segmentation results. It addresses the challenges of resource-intensive traditional segmentation methods and high hardware requirements, offering an efficient solution for skin disease image segmentation tasks.

This study demonstrates the superior performance of the SCSONet model through optimization of parameters and floating-point operations (FLOPs), showcasing its strong generalizability and adaptability compared to other advanced models, while significantly reducing network parameters and computational costs. SCSONet achieves competitive segmentation performance with only 0.149 M parameters and 0.056GFLOPs, making it, to our knowledge, the first model to operate under such low computational load. Notably, SCSONet's lightweight design allows it to be trained with just 0.6 GB of VRAM, a breakthrough feature that not only reduces the dependence on high-performance computing resources but also offers a new solution for medical image segmentation tasks in resource-limited environments. This design focus underscores the innovativeness and practical application value of our model, particularly in advancing mobile health technology and remote medical services.

While SCSONet exhibits a notable reduction in parameters and computational efficiency, it still has a gap compared to EGEUnet in terms of parameter quantity. Additionally, the limited datasets used for experiments and the model's generalizability are areas for further inquiry. Additionally, during multiple training sessions, there were occasional instances of lower accuracy. This indicates that the model may not consistently achieve the expected high precision under certain specific datasets or training conditions, suggesting a sensitivity to training data or a deficiency in the optimization strategy under specific conditions. Although these instances are rare, they must be taken seriously as they could affect the model's reliability and robustness in practical applications.

Future research should focus on extending the lightweight architecture to additional semantic segmentation tasks, alongside a thorough examination of its integration with hardware devices for enhanced performance. Investigating advanced training techniques and structural adjustments to the model will be crucial for augmenting its adaptability and consistency across diverse training scenarios. The ultimate objective is to refine segmentation efficiency without compromising accuracy, thereby rendering the model more effective for assisted diagnostics within medical image analysis. This approach aims to strike a balance between computational efficiency and diagnostic precision, facilitating broader application in real-world clinical settings.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://challenge.isic-archive.com/data/#2017.

Author contributions

HC: Writing–review and editing, Writing–original draft, Software, Methodology, Conceptualization. ZL: Writing–review and editing, Writing–original draft, Software. XH: Writing–review and editing, Writing–original draft, Conceptualization. ZP: Writing–review and editing, Writing–original draft, Validation. YD: Writing–review and editing, Writing–original draft. LT: Formal Analysis, Writing–review and editing, Writing–original draft, Investigation, Conceptualization. LY: Writing–review and editing, Writing–original draft, Investigation, Data curation, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Development Fund of Key Laboratory of Chongqing University Cancer Hospital (cquchkfjj005).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA: A Cancer J Clinicians* (2024) 74:12–49. doi:10.3322/caac.21820

2. Wolf AM, Oeffinger KC, Shih TY-C, Walter LC, Church TR, Fontham ET, et al. Screening for lung cancer: 2023 guideline update from the american cancer society. *CA: A Cancer J Clinicians* (2023) 74:50–81. doi:10.3322/caac.21811

3. Carli P, De Giorgi V, Soyer H, Stante M, Mannone F, Giannotti B. Dermatoscopy in the diagnosis of pigmented skin lesions: a new semiology for the dermatologist. *J Eur Acad Dermatol Venereol* (2000) 14:353–69. doi:10.1046/j.1468-3083.2000.00122.x

4. Dinnes J, Deeks JJ, Chuchu N, di Ruffano LF, Matin RN, Thomson DR, et al. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database Syst Rev* (2018) 12. doi:10.1002/14651858.cd011902.pub2

5. Zhu S, Gao R. A novel generalized gradient vector flow snake model using minimal surface and component-normalized method for medical image segmentation. *Biomed Signal Process Control* (2016) 26:1–10. doi:10.1016/j.bspc.2015.12.004

6. Gupta D, Anand RS. A hybrid edge-based segmentation approach for ultrasound medical images. *Biomed Signal Process Control* (2017) 31:116–26. doi:10.1016/j.bspc.2016.06.012

7. Fraz MM, Jahangir W, Zahid S, Hamayun MM, Barman SA. Multiscale segmentation of exudates in retinal images using contextual cues and ensemble classification. *Biomed Signal Process Control* (2017) 35:50–62. doi:10.1016/j.bspc.2017.02.012

8. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. *J Med Syst* (2018) 42: 226–13. doi:10.1007/s10916-018-1088-1

9. Shelhamer E, Long J, Darrell T, et al. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* (2017) 39:640–51. doi:10.1109/tpami.2016.2572683

10. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference Proceedings, Part III 18; October 5-9, 2015; Munich, Germany. Springer (2015). p. 234–41.

11. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale (2020). Available at: https://arxiv.org/abs/2010.11929. (Accessed December 17, 2023).

12. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications (2017). Available at: https://arxiv.org/abs/1704.04861. (Accessed November 4, 2023).

13. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 18 2018 to June 23 2018; Salt Lake City, UT, USA (2018). p. 4510–20.

14. Koonce B, Koonce B. Mobilenetv3. In: *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization* (2021). p. 125–44.

15. Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer (2021). Available at: https://arxiv.org/abs/2110.02178. (Accessed January 3, 2024).

16. 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; 4-9 February 2017; San Francisco, California, USA (2017).

17. Singh VK, Kalafi EY, Wang S, Benjamin A, Asideu M, Kumar V, et al. Prior wavelet knowledge for multi-modal medical image segmentation using a lightweight neural network with attention guided features. *Expert Syst Appl* (2022) 209:118166. doi:10.1016/j.eswa.2022.118166

18. Valanarasu JMJ, Patel VM. Unext: mlp-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; September 18-22, 2022; Singapore. Springer (2022). p. 23–33.

19. Ruan J, Xiang S, Xie M, Liu T, Fu Y. Malunet: a multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE); 6-8 December 2022; Las Vegas, Nevada, USA (2022). p. 1150–6.

20. Li C, Zhou A, Yao A. Omni-dimensional dynamic convolution (2022). Available at: https://arxiv.org/abs/2209.07947. (Accessed November 2, 2023).

21. Li J, Wen Y, He L. Scconv: spatial and channel reconstruction convolution for feature redundancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18 2022 to June 24 2022; New Orleans, LA, USA (2023). p. 6153–62.

22. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE); June 4 to June 9, 2023; Rhodes Island, Greece (2023). p. 1–5.

23. Chen J, Kao S-h., He H, Zhuo W, Wen S, Lee C-H, et al. Run, don't walk: chasing higher flops for faster neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 17 2023 to June 24 2023; Vancouver, BC, Canada (2023). p. 12021–31.

24. Ahmed SF, Rahman FS, Tabassum T, Bhuiyan MTI. 3d u-net: fully convolutional neural network for automatic brain tumor segmentation. In: 2019 22nd International Conference on Computer and Information Technology (ICCIT) (IEEE); 18-20 December 2019; Dhaka, Bangladesh (2019). p. 1–6.

25. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV) (Ieee); 25-28 October 2016; Stanford, California, USA (2016). p. 565–71.

26. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018 Proceedings 4; September 20, 2018; Granada, Spain. Springer (2018). p. 3–11.

27. Liu Y, Mu F, Shi Y, Chen X. Sf-net: a multi-task model for brain tumor segmentation in multimodal mri via image fusion. *IEEE Signal Process. Lett* (2022) 29:1799–803. doi:10.1109/LSP.2022.3198594

28. Zhu Z, Yao Z, Zheng X, Qi G, Li Y, Mazur N, et al. Drug–target affinity prediction method based on multi-scale information interaction and graph optimization. *Comput Biol Med* (2023) 167:107621. doi:10.1016/j.compbiomed.2023.107621

29. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. *Inf Fusion* (2023) 91: 376–87. doi:10.1016/j.inffus.2022.10.022

30. Li Y, Wang Z, Yin L, Zhu Z, Qi G, Liu Y. X-net: a dual encoding–decoding method in medical image segmentation. *Vis Comp* (2021) 1–11.

31. Liu Y, Shi Y, Mu F, Cheng J, Chen X. Glioma segmentation-oriented multi-modal mr image fusion with adversarial learning. *IEEE/CAA J Automatica Sinica* (2022) 9: 1528–31. doi:10.1109/JAS.2022.105770

32. He X, Qi G, Zhu Z, Li Y, Cong B, Bai L. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. *Simulation Model Pract Theor* (2023) 126:102769. doi:10.1016/j.simpat.2023.102769

33. Zhuang Y, Liu H, Song E, Xu X, Liao Y, Ye G, et al. A 3-d anatomy-guided self-training segmentation framework for unpaired cross-modality medical image segmentation. *IEEE Trans Radiat Plasma Med Sci* (2024) 8:33–52. doi:10.1109/TRPMS.2023.3332619

34. Ruiping Y, Kun L, Shaohua X, Jian Y, Zhen Z. Vit-upernet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation. *Complex Intell Syst* (2024) 1–13. doi:10.1007/s40747-024-01359-6

35. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* (2012) 25.

36. Berseth M. Isic 2017-skin lesion analysis towards melanoma detection (2017). Available at: https://arxiv.org/abs/1703.00523. (Accessed January 9, 2024).

37. Ruan J, Xie M, Gao J, Liu T, Fu Y. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; October 8-12, 2023; Vancouver, BC, Canada. Springer (2023). p. 481–90.

38. Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference Proceedings, Part I 24; September 27–October 1, 2021; Strasbourg, France. Springer (2021). p. 14–24.

39. Le PT, Pham B-T, Chang C-C, Hsu Y-C, Tai T-C, Li Y-H, et al. Anti-aliasing attention u-net model for skin lesion segmentation. *Diagnostics* (2023) 13:1460. doi:10.3390/diagnostics13081460

40. Dayananda C, Yamanakkanavar N, Nguyen T, Lee B. Amcc-net: an asymmetric multi-cross convolution for skin lesion segmentation on dermoscopic images. *Eng Appl Artif Intelligence* (2023) 122:106154. doi:10.1016/j.engappai.2023.106154

41. Jiang X, Jiang J, Wang B, Yu J, Wang J. Seacu-net: attentive convlstm u-net with squeeze-and-excitation layer for skin lesion segmentation. *Comp Methods Programs Biomed* (2022) 225:107076. doi:10.1016/j.cmpb.2022.107076

42. Wu H, Chen S, Chen G, Wang W, Lei B, Wen Z. Fat-net: feature adaptive transformers for automated skin lesion segmentation. *Med image Anal* (2022) 76: 102327. doi:10.1016/j.media.2021.102327

43. Fan R, Wang Z, Zhu Q. Egfnet: efficient guided feature fusion network for skin cancer lesion segmentation. In: 2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI); October 26-28, 2024; Tipasa, ALGERIA (2022). p. 95–9.

44. Zhao C, Lv W, Zhang X, Yu Z, Wang S. Mms-net: multi-level multi-scale feature extraction network for medical image segmentation. *Biomed Signal Process Control* (2023) 86:105330. doi:10.1016/j.bspc.2023.105330

45. Wang J, Huang G, Zhong G, Yuan X, Pun C-M, Deng J. Qgd-net: a lightweight model utilizing pixels of affinity in feature layer for dermoscopic lesion segmentation. *IEEE J Biomed Health Inform* (2023) 27:5982–93. doi:10.1109/jbhi.2023.3320953

46. Ma Q, Mao K, Wang G, Xu L, Zhao Y. Lcaunet: a skin lesion segmentation network with enhanced edge and body fusion (2023). Available at: https://arxiv.org/pdf/2305.00837.pdf. (Accessed October 28, 2023).

TYPE Original Research
PUBLISHED 26 March 2024
DOI 10.3389/fphy.2024.1356248

OPEN ACCESS

EDITED BY
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China

REVIEWED BY
Yonghang Tai,
Yunnan Normal University, China
Taisong Jin,
Xiamen University, China
Jie Liu,
North China University of Technology, China

*CORRESPONDENCE
SuWei Zhai,
✉ suwei_zhai@163.com

RECEIVED 15 December 2023
ACCEPTED 15 January 2024
PUBLISHED 26 March 2024

CITATION
Feng Y, Luo E, Lu H and Zhai S (2024), Cross-modality feature fusion for night pedestrian detection.
Front. Phys. 12:1356248.
doi: 10.3389/fphy.2024.1356248

Cross-modality feature fusion for night pedestrian detection

Yong Feng, Enbo Luo, Hai Lu and SuWei Zhai*

Electric Power Research Institute, Yunnan Power Grid Corporation, Kunming, China

Night pedestrian detection with visible image only suffers from the dilemma of high miss rate due to poor illumination conditions. Cross-modality fusion can ameliorate this dilemma by providing complementary information to each other through infrared and visible images. In this paper, we propose a cross-modal fusion framework based on YOLOv5, which is aimed at addressing the challenges of night pedestrian detection under low-light conditions. The framework employs a dual-stream architecture that processes visible images and infrared images separately. Through the Cross-Modal Feature Rectification Module (CMFRM), visible and infrared features are finely tuned on a granular level, leveraging their spatial correlations to focus on complementary information and substantially reduce uncertainty and noise from different modalities. Additionally, we have introduced a two-stage Feature Fusion Module (FFM), with the first stage introducing a cross-attention mechanism for cross-modal global reasoning, and the second stage using a mixed channel embedding to produce enhanced feature outputs. Moreover, our method involves multi-dimensional interaction, not only correcting feature maps in terms of channel and spatial dimensions but also applying cross-attention at the sequence processing level, which is critical for the effective generalization of cross-modal feature combinations. In summary, our research significantly enhances the accuracy and robustness of nighttime pedestrian detection, offering new perspectives and technical pathways for visual information processing in low-light environments.

KEYWORDS

pedestrian detection, YOLOv5, vision transformer, CNNs, feature fusion

1 Introduction

Pedestrians are a vital element in traffic scenarios, and the ability to detect pedestrians quickly and accurately has increasingly become a critical research topic in the field of computer vision. Pedestrian detection plays an essential role in various practical applications, such as autonomous driving perception systems [1–3] and intelligent security monitoring systems [4–6]. Additionally, pedestrian detection serves as the foundational task for downstream tasks like pedestrian tracking [7–9], action recognition and prediction [10–12], with its accuracy directly impacting the performance of these tasks. With the significant advancements in convolutional neural networks (CNNs), pedestrian detection models [13–16] have been continually updated and iterated, bringing forth models with outstanding performance. However, most pedestrian detection models are trained on single-modality, well-illuminated visible light datasets [17–19]. When faced with low-light conditions such as at night, their performance significantly declines due to excessive noise and decreased discriminability [4, 20]. Pedestrian detection using only nighttime visible light images is particularly challenging because the data modality itself lacks a valid target area. Therefore, an increasing amount of research is focusing on cross-modality fusion learning, such as the fusion detection of visible and infrared images [21–26].

Infrared vision sensors operate on the principle of thermal imaging, distinguishing pedestrians from the background by differences in thermal radiation. Infrared imagery is robust against interference and is not easily affected by adverse environmental conditions [27, 28]. Even at night, infrared images can reveal the shape of pedestrians, effectively compensating for the vulnerability of visible light images to lighting conditions. However, infrared images also have drawbacks, such as lower resolution and a lack of texture information. On the other hand, visible light images provide rich detail and texture information [22]. Therefore, cross-modal fusion aims to extract complementary information between these two modalities, enhancing the flow of information between them and improving the perceptibility and robustness of detection algorithms. In the field of image fusion, a lot of work [29] has been carried out on the effective fusion of infrared images and visible light images.

In the field of pedestrian detection that fuses visible and infrared imaging, many approaches rely solely on Convolutional Neural Networks (CNN) to extract deep features [21, 23, 25, 26], with artificially designed complex fusion mechanisms to integrate features from different modalities. Extensive research has demonstrated the powerful representational capabilities of CNNs for expressing visual features in single-modality scenarios [30–32]. However, due to the limited receptive field, CNNs, while adept at capturing local information, exhibit weaker capabilities in capturing global texture information across modalities in fusion tasks. Transformer [33, 34] is equipped with self-attention mechanisms, possess a global receptive field and excel at learning long-range dependencies. Therefore, combining CNNs with transformers for cross-modality nighttime pedestrian detection can leverage the strengths of both, resulting in complementary advantages and enhanced detection performance.

Recently, vision transformers [33, 35–37] have been processing inputs as sequences and have demonstrated the capability to capture long-range correlations, offering a promising avenue towards a unified framework for multi-modal tasks. However, it remains to be clarified whether vision transformers can bring potential improvements to vis-inf pedestrian detection compared to existing multi-modal fusion modules [38–40] based on Convolutional Neural Networks (CNNs). Crucially, while some earlier studies have employed a simplistic global multi-modal interaction strategy, such an approach has not been universally applicable across various sensing data combinations [41–43]. We posit that in vis-inf pedestrian detection, which involves a variety of supplementary information and uncertainties, a comprehensive cross-modal interaction should be implemented to fully leverage the potential of cross-modal complementary features.

To address the challenges in vis-inf nighttime pedestrian detection, we propose an interactive cross-modal fusion framework based on yolov5, named FRFPD. This framework aims to enhance the performance of detection algorithms through efficient information fusion. FRFPD is constructed as a dual-stream architecture, specifically handling visible light (VIS) and infrared (Inf) data streams. On this foundation, we have designed feature interaction and fusion modules to optimize model performance: The Cross-Modal Feature Rectification Module (CMFRM) fine-tunes VIS and Inf features at a granular level, utilizing their spatial correlations to enhance the model's focus on complementary information and effectively reduce the uncertainty and noise from different modalities. This process precisely handles the complexity of multi-source data, paving the way for more effective feature extraction and interaction. Moreover, the Feature Fusion Module (FFM) [41] is structured in two stages, ensuring ample information exchange before feature fusion on a global scale. In the first stage, we introduce a cross-attention mechanism for cross-modal global reasoning, propelled by a wide receptive field facilitated by the self-attention mechanism. In the second stage, a mixed channel embedding is employed to generate enhanced feature outputs. In essence, the interaction strategy we introduce is multidimensional: within the CMFRM module, we correct feature maps on a spatial dimension; while in the FFM module, it apply a cross-modal attention mechanism for feature fusion across the global channel dimension. These approaches are vital for the effective generalization of cross-modal feature combinations, enhancing the model's capability to process information from diverse sensory modalities. Our contributions are summarized as follows:

(1) A dual-stream architecture is proposed in the FRFPD framework, leveraging YOLOv5, to handle visible light (VIS) and infrared (INF) data streams separately, tailored for addressing low-light challenges in nighttime pedestrian detection.
(2) The Cross-Modal Feature Rectification Module (CMFRM) is introduced to fine-tune visible and infrared features, exploiting their spatial correlations to enhance focus on complementary information, significantly reducing uncertainty and noise from different modalities. NF.
(3) An advanced Feature Fusion Module (FFM) developed in [41] is introduced, in two stages to promote ample information exchange and utilize a mixed channel embedding for generating enhanced feature outputs, improving detection capabilities.

2 Related works

The widespread application of Transformers in the field of Natural Language Processing (NLP) has proven their excellence and convenience in handling sequential data, which has also made them popular for visual tasks.

2.1 Vision transformer

The widespread application of Transformers in the field of Natural Language Processing (NLP) has proven their excellence and convenience in handling sequential data, which has also made them popular for visual tasks [35, 36, 44]; [45, 46]. ViT [35] addresses the high computational cost issue of Transformers in traditional visual tasks by flattening images into a series of pixel blocks (patches), transforming image processing tasks into a form similar to the word sequence processing in NLP. DeiT [47] further proposes a convolution-free Transformer structure, introducing a teacher-student strategy through distillation tokens, with training conducted solely on ImageNet. Moreover, the positional encoding feature of Transformers is used to capture the order information of sequence data, which can be either fixed or learnable [48].

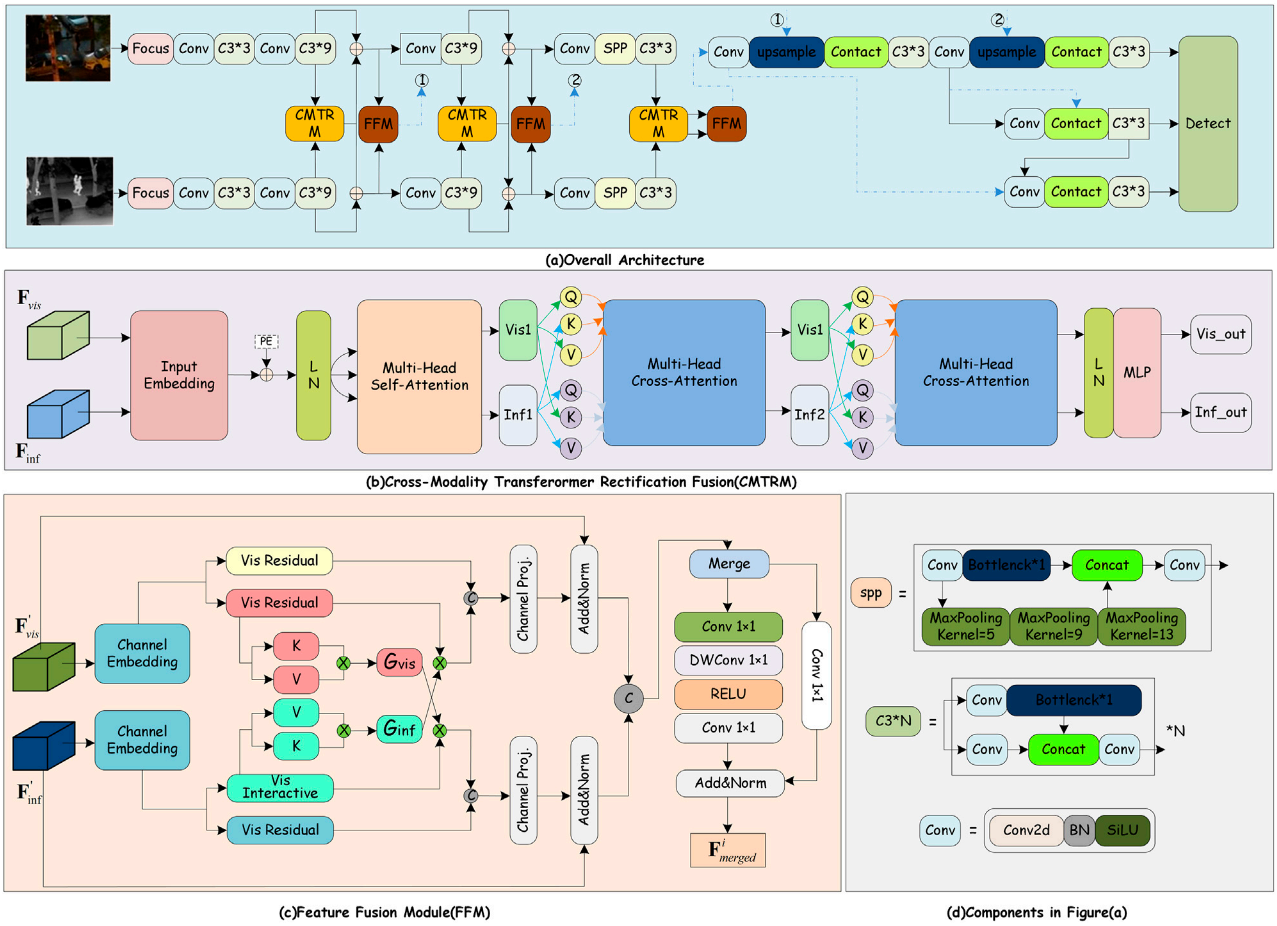
In the field of computer vision, Visual Transformer (VT) have demonstrated significant capabilities across various tasks such as image Fusion [49, 50]), pedestrian detection [51], particularly excelling in multispectral detection tasks [52–55] where they can focus on important features scattered across different spectral bands. Their self-attention mechanism's ability to model long-range dependencies and capture global context is especially valuable. Unlike convolutional neural networks [26, 56–58], VT operate on sequences of image patches (tokens) and are adept at learning to concentrate on the most informative parts of the input, making them inherently suited for multispectral detection where significant features may be sparsely distributed across spectral bands. However, the application of VT in multispectral detection, especially under challenging lighting conditions, remains a developing field. Our work is inspired by the intrinsic advantages of VT to tackle unique challenges in low-light multispectral scenarios. We have introduced a novel VT-based framework, specifically designed for this purpose, that incorporates modules sensitive to the nuances of multispectral data. Our proposed Cross-Modal Feature Rectification Module (CMFRM) expands the concept of VT by integrating cross-modal learning directly into the transformer architecture, serializing tokens along the spatial dimension, thereby enhancing the model's ability to perform fine-grained feature adjustment. This is critical for aligning features across different modalities, particularly when contending with varying levels of illumination and noise inherent in low-light conditions.

2.2 Multispectral pedestrian detection

The field of pedestrian detection has seen the emergence of numerous outstanding studies, including early traditional detection methods [59, 60] and the surge of CNN-based detection technologies [61–64] that came with the rapid development of Convolutional Neural Networks (CNN). However, the majority of research is still focused on single-modality visible light images. In nocturnal environments, relying solely on visible light images for pedestrian detection often fails to achieve satisfactory results, mainly because conventional visible light cameras perform poorly in night-time imaging, with target areas not being distinct and substantial noise interference. For this reason, it becomes extremely difficult for models like CNNs to extract effective features from nighttime visible light images. As research has deepened, infrared imagery, with its unique advantages in night-time settings, has started to be used to complement the shortcomings of visible light images. This has attracted increasing attention from researchers and has spurred the advancement and exploration of multispectral pedestrian detection technologies, especially those based on CNN approaches.

In the field of multispectral detection, fusion algorithms play a crucial role. The AR-CNN [65] model introduces an end-to-end region alignment algorithm, which addresses the subtle misalignments caused by positional offsets between multimodalities. This fusion approach reweights features to prioritize more reliable characteristics and suppress ineffective ones. Meanwhile, the CIAN [26] model leverages the interactive properties of multispectral input sources, proposing a cross-channel interactive attention network. This network extracts global features from each channel of the two modalities and recalibrates the channel responses of intermediate feature maps using an attention mechanism by computing the inter-channel correlation. In existing multispectral detection research, models like AR-CNN and CIAN offer solutions for minor misalignments between modalities and feature recalibration; however, these methods still show limitations in complex scenarios under low-light conditions, such as night-time pedestrian detection. These limitations manifest in two aspects: firstly, feature information loss due to insufficient lighting under low-light conditions cannot be compensated for by simply reweighting features; secondly, despite the CIAN model employing an interactive attention mechanism, more efficient strategies for information exchange and fusion are needed to handle the complex interactions between different modalities. CFT [66] proposed a fusion algorithm that combines transformer and CNN, which can learn remote dependencies and extract global context information. Self-attention can fuse features within and between modes. It is a relatively novel method recently, but this model uses traditional transformer, which has the problems of positional encoding and multi-head attention mismatch cross-modality fusion. ProbEn [67] research primarily focuses on the issue of multimodal object detection, with a particular emphasis on addressing the challenges of object detection in low-light conditions. It introduces the ProbEn probabilistic ensemble technique to effectively fuse object detection results from different sensors, thereby significantly enhancing the performance of multimodal object detection. UGC [68] is dedicated to addressing crucial challenges in multispectral pedestrian detection, encompassing issues such as image calibration and disparities between different modalities. The authors introduce a novel approach that aims to enhance pedestrian detection performance by incorporating Region of Interest (RoI) uncertainty and predictive uncertainty into the feature fusion and modality alignment processes.

To overcome these limitations, we propose the FRFPD framework, central to which are the Cross-Modal Feature Rectification Module (CMFRM) and the Feature Fusion Module (FFM). The CMFRM is motivated by the need to serialize tokens in the spatial dimension for fine-grained feature adjustment, aligning features within the visible and infrared modalities. Its design aims to finely tune features across modalities by exploiting their spatial correlations to amplify complementary information, thereby significantly reducing uncertainty and noise in low-light conditions. This approach is crucial for enhancing the accuracy and robustness of detection under varied lighting conditions. Concurrently, the FFM addresses the challenge of integrating diverse modalities effectively. It serializes tokens globally in the channel dimension, first performing global reasoning between modalities through a cross-attention mechanism, then refining the feature output with hybrid channel embedding. This strategy is driven by the need to provide not only an in-depth exchange of information but also a more nuanced enhancement of channel responses than the CIAN model. The motivation behind FFM is to improve the overall quality of feature fusion, enhancing the detection capabilities in complex scenarios. The FRFPD



(a)Overall Architecture

(b)Cross-Modality Transferormer Rectification Fusion(CMTRM)

(c)Feature Fusion Module(FFM)

(d)Components in Figure(a)

FIGURE 1
The Network structure of our proposals. **(A)** shows our overall network architecture, which adopts a novel combination of CNN and transformer. The deep features of visible and infrared images are extracted by two-stream CNN, and the proposed CMFRM module is used to leverag the features from one modality to rectify the features of the other modality. Feature Fusion Module (FFM) operates through a bifurcated process, as illustrated in **(C)** an initial stage of global information exchange followed by a stage of comprehensive global feature fusion. This structure is designed to facilitate extensive information interchange preceding the fusion of features at a global level. In addition, **(D)** shows the structure of the components in **(A)**.

framework sets a new performance benchmark for cross-modal feature fusion through its multi-dimensional interaction strategy, correcting feature maps on the channel and spatial dimensions, and implementing cross-attention at the sequence processing level.

3 Proposed method

3.1 Overview

Among the numerous target detection CNN models, YOLOv5 [69] is a highly reliable algorithm with fast recognition speed, which is easier to deploy and train. It is also one of the most popular detection frameworks currently and has a wide range of applications. Therefore, in this paper, we choose YOLOv5 to extract deep features and extend the transformer fusion algorithm to a dual-stream architecture. The backbone of YOLOv5 is modified from a single-stream structure to a dual-stream structure to separately extract deep features of the input visible light and infrared images. The rectification module, called Cross-Modal Feature Rectification Module (CMFRM), is implemented three times in the backbone. CMFRM is corrected one feature against another, and *vice versa*. In this way, the features of both modalities can be corrected. Additionally, as illustrated in Figure 1B, we introduced a Feature Fusion Module (FFM) [41] that merges features belonging to the same level into a single feature map. Then, a detection head is used to predict the final pedestrian positions. Our proposed network framework is illustrated in Figure 1.

3.2 Cross-modality feature rectification module

In this paper, we explore the complementarity of information from different sensors [8], [9], noting that while this information is valuable, it is often affected by noise. To address this issue, we introduce a novel Cross-Modal Feature Rectification Module (CMFRM) in Figure 1B, which is capable of performing precise feature correction at each stage of feature extraction on parallel data streams. Utilizing Transformer technology for spatial feature correction, the CMFRM provides a

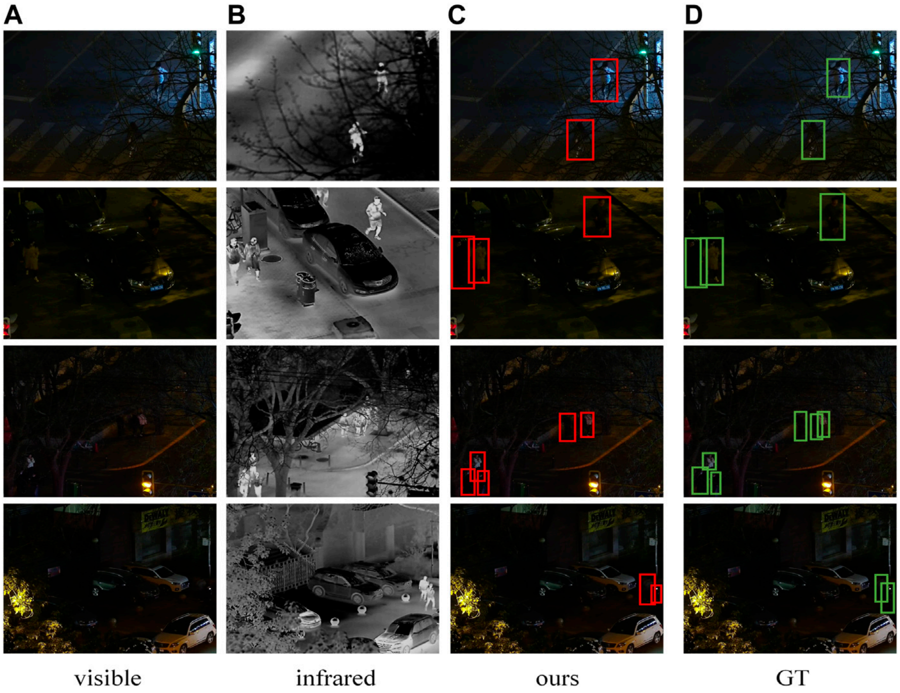


FIGURE 2
The visualization of the detection results, subfigure **(A)** shows the input visible lr images, subfigure**(B)** is the corresponding infrared images, subfigure **(C)** is the prediction result of our model, and subfigure **(D)** is the ground truth. These images are selected from the dataset listed at https://soonminhwang.github.io/rgbt-ped-detection/

granular correction mechanism. This not only effectively handles noise and uncertainty across different sensory modalities but also enhances the extraction and interaction of multimodal features, thereby improving the overall performance of the system.

In a two-stream structure, we extract features from visible and infrared images independently through Convolutional Neural Networks (CNN), obtaining visible feature and infrared feature, respectively. Both feature sets have the shape (B, C, H, W), where B is the batch size, C is the number of channels, and H and W are the dimensions of the spatial size. To adapt these features for the transformer, we flatten them into the shape (B, N, C), while proceeding along the spatial dimensions. where N is the number of tokens, given by $N = H \times W$. This step is a crucial phase in the transition of CNN features to transformer-based CMFRM module.

$$\text{flat}_{vis} = F_{vis} \cdot \text{view}(B, C, -1) \tag{1}$$

$$\text{flat}_{inf} = F_{inf} \cdot \text{view}(B, C, -1) \tag{2}$$

$$\text{flat}_{\text{cat}} = \text{concat}\left(\left(flat_{vis}, flat_{\text{inf}}\right), \text{dim} = 2\right) \tag{3}$$

$$Z = \text{flat}_{\text{cat}}.\text{permute}(0, 2, 1) \tag{4}$$

where F_{vis} and F_{inf} represent the visible and infrared features from the CNN, respectively. The `view` function reshapes the tensor of specified shape without changing its data, and `concat` concatenates the given tensors along the specified dimension. The `permute` function outputs a tensor after permuting the dimensions of the input tensor. Thus, in Eq 4, the shape of Z is $(B, 2N, C)$.

Positional embeddings enable the model to discern spatial relationships between different tokens during training. After positional embedding, the input sequence Z is then projected onto three weight matrices to compute a set of queries, keys, and values (Q, K, $andV$), expressed as $Q = ZW^Q$, $K = ZW^K$, $V = ZW^V$. In this context, the weight matrices are defined as $W^Q \in \mathbb{R}^{C\times D_Q}$, $W^K \in \mathbb{R}^{C\times D_K}$, and $W^V \in \mathbb{R}^{C\times D_V}$. Furthermore, the dimensions D_Q, D_K, and D_V are equivalent in our transformer model, such that $D_Q = D_K = D_V = C$. The Multi-head Self-Attention layer computes the attention weights by calculating the scaled dot products between Q and K. These weights are then applied to V to infer the refined output $\hat{Z}$.

$$\hat{Z} = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_K}}\right)V \tag{5}$$

However, multimodal data is distributed across different spatial domains, and relying solely on self-attention is insufficient for fully exploiting the mixed modality information, which may result in inadequate rectification. Based on the principle of self-attention, we speculate that exchanging the “values” and “keys” between different modalities might better enhance the vital information and facilitate the flow of complementary information. Building on these considerations, we have extended the traditional multi-head attention based on a cascading strategy by incorporating two instances of Cross-Attention (CA), as shown in Figure 1B. Additionally, the process of information exchange during the two instances of Cross-Attention can be represented by Eqs. 6–9.

$$CA^1_{vis}\left(Q_{vis}, K_{inf}, V_{vis}\right) = \text{softmax}\left(\frac{Q_{vis}K^T_{inf}}{\sqrt{d_k}}\right)V_{vis} \tag{6}$$

$$CA^1_{inf}\left(Q_{inf}, K_{vis}, V_{inf}\right) = \text{softmax}\left(\frac{Q_{inf}K^T_{vis}}{\sqrt{d_k}}\right)V_{inf} \tag{7}$$

and

$$CA_{vis}^{2}\left(Q_{vis}, K_{vis}, V_{inf}\right) = \mathrm{softmax}\left(\frac{Q_{vis}K_{vis}^{T}}{\sqrt{d_k}}\right)V_{inf} \tag{8}$$

$$CA_{inf}^{2}\left(Q_{inf}, K_{inf}, V_{vis}\right) = \mathrm{softmax}\left(\frac{Q_{inf}K_{inf}^{T}}{\sqrt{d_k}}\right)V_{vis} \tag{9}$$

where vis, inf represent visible token and infrared token from $\hat{Z}$ respectively. After processing through two cascaded multi-head cross-attention layers, the visible and infrared features are subjected to Layer Normalization (LN) and Multi-Layer Perceptron (MLP), ultimately producing two output features, $\tilde{\mathbf{F}}_{vis}$ and $\tilde{\mathbf{F}}_{inf}$.

3.3 Two-stage feature fusion module

After obtaining the feature mappings from each layer, a two-stage feature fusion module (Feature Fusion Module, FFM) [41] is introduced to enhance the interaction and integration of global information. As illustrated in Figure 1C, in the first stage, the two branches are kept separate, and a cross-attention mechanism is designed to facilitate the global exchange of information between the two branches. In the stage 2, the concatenated features are transformed back to the original scale through a mixed channel embedding.

Global Information exchange stage. We first flatten the input feature of size $\tilde{\mathbf{F}}_{vis}$ and $\tilde{\mathbf{F}}_{inf} \in \mathbb{R}^{H\times W\times C}$ into $R^{N\times C}$ along with channel dimension, where $N = H \times W$, and C is the number of tokens, Then, through linear embedding, we generate two vectors of the same size $R^{N\times C}$, named the residual vector X_{res} and the interactive vector X_{inter}. Building upon this, we propose an efficient cross-attention mechanism that applies to these two interactive vectors from different modal pathways, achieving comprehensive information exchange across modalities. This mechanism offers complementary interactions from a sequence-to-sequence perspective, surpassing the rectification-based interactions from the feature map perspective in CMFRM.

Our cross-attention mechanism, designed for improved cross-modal feature fusion, is an adaptation of the conventional self-attention mechanism [33]. The traditional method encodes inputs into Queries (Q), Keys (K), and Values (V), computing a global attention map via QK^T. This results in a computationally expensive $N \times N$ matrix. Alternatively [70], proposes using a global context vector $G = K^TV$, reducing the size to $C_{\mathrm{head}} \times C_{\mathrm{head}}$. Our approach builds on this by embedding interactive vectors into K and V for each head, with both matrices sized $N \times C_{\mathrm{head}}$. The final output is a product of these interactive vectors and the context vector from an alternate modality, constituting the cross-attention process.

$$\begin{aligned} G_{vis} &= \hat{K}_{vis}^{T}\hat{V}_{vis} \\ G_{inf} &= \hat{K}_{inf}^{T}\hat{V}_{inf} \end{aligned} \tag{10}$$

$$\begin{aligned} U_{vis} &= X_{vis}^{\mathrm{inter}}\mathrm{Softmax}\left(G_{inf}\right) \\ U_{inf} &= X_{inf}^{\mathrm{inter}}\mathrm{Softmax}\left(G_{vis}\right) \end{aligned} \tag{11}$$

Note that G denotes the global context vector, while U indicates the attended result vector. To realize attention across different representational subspaces, we maintain the multi-head mechanism, where the number of heads corresponds to the number of elements in the transformer backbone. Subsequently, the attended result vector U and the residual vector are concatenated. Finally, we apply a second linear embedding and resize the feature back to $R^{H\times W\times C}$.

Global Feature Fusion Module. In the fusion component of the Feature Fusion Module (FFM), channel-wise integration is performed using 1 × 1 convolution for combining features from dual pathways. Considering the necessity of spatial context for Vis-Inf pedestrain detection, we adopt a strategy influenced by Mix-FFN [71] and ConvMLP [72], incorporating a depth-wise 3 × 3 convolution (DW Conv) to form a skip connection architecture. This approach facilitates the consolidation of the concatenated feature dimensions $R^{H\times W\times 2C}$ into the decoder output dimension $R^{H\times W\times C}$.

4 Experiments

In this section, we first introduce two multispectral datasets, KAIST [73] and LLVIP [22]. The KAIST dataset compiles data from day and night autonomous driving scenarios, while the LLVIP dataset is composed of night-time surveillance scenarios. Given our focus on nighttime pedestrian detection, we exclusively selected the nighttime subset of the KAIST dataset. Subsequently, we delve into some specifics of the model training phase. The evaluation metrics for pedestrian detection diverge slightly from those of traditional object detection, hence we will clarify the evaluation metrics utilized in this study. We benchmark our results against state-of-the-art methods and conduct ablation studies to assess the effectiveness of our proposed module. Lastly, the visualization of our proposals is provided to facilitate an intuitive understanding of their impact. At last, we provide a visualization of the predicted results as shown in Figure 2.

4.1 Dataset

KAIST. The KAIST dataset [73], introduced at CVPR2015, consists of 95k aligned pairs of visible and infrared images and has been extensively utilized. All annotations are manually labeled, including 1,182 pedestrian instances. Due to biased annotations in the original training set, this study employs the sanitized version [23]. The sanitized KAIST provides 7,601 training images with at least one valid pedestrian instance, filtered and sampled from the original training videos. There are 2,846 pairs for night training and 4,755 pairs for day training. The test set comprises 2,252 image pairs, with 797 for night and 1,455 for day. Test annotations from the improved version [31], which corrects the initial annotations, are used. The resolution of training and test images is 640 × 512.

LLVIP. LLVIP [22] is a nighttime pedestrian dataset for surveillance scenarios, presented at ICCV2021. It includes 15,488 strictly aligned visible-infrared image pairs, featuring numerous pedestrians and cyclists from diverse street locations between 6 and 10 p.m. [22]. The original resolution of the images is 1280 × 1024, but to reduce computational demands, we scale down the images by half to 640 × 512 in this paper.

TABLE 1 Results on KAIST night dataset and the results in bold indicate the optimal.

Methods	Data modality	LAMR (%)	AP50
Yolov5 [69]	Visible	63.65	43.95%
Yolov5 [69]	Infrared	14.73	77.51%
MLF-CNN [74]	Visible + Infrared	25.65	67.60%
IATDNN [75]	Visible + Infrared	26.88	67.02%
CWF-CNN [76]	Visible + Infrared	30.82	64.59%
L-SSD [77]	Visible + Infrared	35.38	48.77%
MSDS-RCNN [23]	Visible + Infrared	13.73	-
CS-RCNN [78]	Visible + Infrared	11.86	-
CIAN [26]	Visible + Infrared	11.13	-
MBNet [79]	Visible + Infrared	10.98	-
UGC [68]	Visible + Infrared	10.92	-
ProbEn [67]	Visible + Infrared	10.83	-
Our Method	Visible + Infrared	**10.79**	**82.48%**

4.2 Evaluation

Evaluation metrics. The first assessment metric is the Log-Average Miss Rate (LAMR), which is a specialized metric for evaluating the performance of pedestrian detection systems. The relationship between the Miss Rate (MR) and the False Positives Per Image (FPPI) is plotted on a log-log scale, and nine FPPI reference points are selected within the range [10^{-2}, 10^{0}], evenly spaced in the logarithmic space. LAMR is defined as shown in Eq 14.

$$MR = \frac{FN}{TP + FN} \tag{12}$$

$$FPPI = \frac{FP}{imgs\,\mathrm{num}} \tag{13}$$

$$LAMR = \exp\left(\frac{1}{9}\sum_{f}\log\left(MR\underset{FPPI\le f}{\operatorname{argmax}} FPPI\right)\right) \tag{14}$$

where f is within the set {10^{-2}, $10^{-1.75}$, . . . , 10^{0}}, TP represents the number of True Positives, FP is the number of False Positives, and FN denotes the number of False Negatives. Additionally, we utilize AP50 as our second metric, complementing LAMR. In the evaluation process, all detected bounding boxes are matched to ground truth annotations for each image via a greedy algorithm. If the Intersection over Union (IoU) between the detection box and the ground truth exceeds a specified threshold, the detection is considered a True Positive (TP), indicating a successful prediction. Due to the highly non-rigid nature of pedestrians, we adopt the common IoU threshold of 0.5. Thus, AP50 denotes the Average Precision when the IoU threshold is 0.5.

4.3 Comparison of results on KAIST night dataset

We compared our model with the results of state-of-the-art models on the KAIST Night test set, as presented in Table 1. Our model builds upon a two-stream architecture extended from yolov5; hence, we assessed the single-modality detection capabilities of yolov5 with only visible and only infrared images on the same dataset. The task of night-time pedestrian detection using solely visible light images poses a substantial challenge, reflected in a high LAMR of 63.65%. Through the development of effective cross-modality fusion algorithms, such as MSDS-RCNN [23] and CFT [66], the LAMR for night-time pedestrian detection can be significantly decreased, improving detector performance. Furthermore, our proposed method records a LAMR of 10.79% and an AP50 of 82.48%, evidencing the effectiveness and competitive edge of our approach.

4.4 Ablation study

From the previous sections, we have familiarized ourselves with the architecture and proposed modules such as CMFRM, as well as the enhancements in our method. However, the exact quantitative improvements contributed by these modules remain uncertain. Therefore, in this section, we present a succinct and insightful ablation study to address the aforementioned inquiries. Table 2 illustrates that CMFRM has led to a decrease of 1.14% in LAMR and an enhancement of 1.47% in AP50 on the KAIST Night dataset, and a reduction of 0.63% in LAMR on the LLVIP dataset. FFM contributes to a decrease of 0.57% in LAMR and an improvement of 1.18% in AP50 on the KAIST Night dataset, and a reduction of 0.80% in LAMR on the LLVIP dataset. Finally, when compared to the baseline model CFT [66], our comprehensive model CMTF decreases LAMR by 1.38% and enhances AP50 by 3.2% on the KAIST Night dataset, and lowers LAMR by 1.62% on the LLVIP dataset.

TABLE 2 Results of ablation study and the results in bold indicate the optimal.

	Method		KAIST night		LLVIP	
Base	CMFRM	FFM	LAMR (%)	AP50 (%)	LAMR (%)	AP50 (%)
✓			12.71	79.28	5.40	97.50
✓	✓		11.57	80.75	4.77	97.72
✓		✓	12.14	80.46	4.60	97.09
✓	✓	✓	**10.79**	**82.48**	**3.78**	**97.98**

4.5 Conclusion

In this paper, we introduce an interactive cross-modal fusion framework based on YOLOv5, designed to improve the performance of nighttime pedestrian detection algorithms through efficient information fusion. Our framework utilizes a dual-stream architecture to separately handle visible and infrared images, effectively addressing the challenges posed by low-light conditions. Our proposed FRFPD significantly enhance model performance by fine-tuning features across modalities, reducing uncertainty and noise, and focusing on complementary information. These modules also facilitate multi-dimensional feature interaction and rectification, including cross-attention mechanisms at the sequence processing level, which are crucial for the effective generalization of cross-modal feature combinations. Overall, our research not only boosts the performance of nighttime pedestrian detection but also offers new technical solutions and perspectives for visual information processing under low-light conditions.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://multispectral.kaist.ac.kr/pedestrian/data-kaist.

Author contributions

YF: Methodology, Writing–original draft. EL: Writing–review and editing. HL: Writing–review and editing. SZ: Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

Authors YF, EL, HL, and SZ were employed by Yunnan Power Grid Corporation.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Chen L, Lin S, Lu X, Cao D, Wu H, Guo C, et al. Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey. *IEEE Trans Intell Transportation Syst* (2021) 22:3234–46. doi:10.1109/tits.2020.2993926

2. Chen Z, Huang X. Pedestrian detection for autonomous vehicle using multi-spectral cameras. *IEEE Trans Intell Vehicles* (2019) 4:211–9. doi:10.1109/tiv.2019.2904389

3. Hbaieb A, Rezgui J, Chaari L. Pedestrian detection for autonomous driving within cooperative communication system. In: *2019 IEEE wireless communications and networking conference (WCNC)*. IEEE (2019). p. 1–6.

4. Wang X, Chen J, Wang Z, Liu W, Satoh S, Liang C, et al. When pedestrian detection meets nighttime surveillance: a new benchmark. *International Joint Conference on Artificial Intelligence* (2020) 20000:509–515. doi:10.24963/ijcai.2020/71

5. Kulbacki M, Segen J, Wojciechowski S, Wereszczyński K, Nowacki JP, Drabik A, et al. Intelligent video monitoring system with the functionality of online recognition of people?s behavior and interactions between people. In: *Intelligent information and database systems: 10th asian conference, ACIIDS 2018, dong hoi city, vietnam, march 19-21, 2018, proceedings, Part II 10*. Springer (2018). p. 492–501.

6. Rai M, Husain AA, Maity T, Yadav RK, Neves A. Advance intelligent video surveillance system (aivss): a future aspect. *Intell Video Surveill* (2019) 37. doi:10.5772/intechopen.76444

7. Huang L, Zhao X, Huang K. Bridging the gap between detection and tracking: a unified approach. *Proc IEEE/CVF Int Conf Comput Vis* (2019) 3999–4009. doi:10.1109/ICCV.2019.00410

8. Sun Z, Chen J, Chao L, Ruan W, Mukherjee M. A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Trans Circuits Syst Video Technol* (2020) 31:1819–33. doi:10.1109/tcsvt.2020.3009717

9. Stadler D, Beyerer J. Improving multiple pedestrian tracking by track management and occlusion handling. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021). p. 10958–67.

10. Zhang P, Lan C, Zeng W, Xing J, Xue J, Zheng N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020).

11. Liu K, Liu W, Ma H, Tan M, Gan C. A real-time action representation with temporal encoding and deep compression. *IEEE Trans Circuits Syst Video Technol* (2020) 31:647–60. doi:10.1109/tcsvt.2020.2984569

12. Kong Y, Fu Y. Human action recognition and prediction: a survey. *Int J Comput Vis* (2022) 130:1366–401. doi:10.1007/s11263-022-01594-9

13. Huang X, Ge Z, Jie Z, Yoshie O. Nms by representative region: towards crowded pedestrian detection by proposal pairing. *Proc IEEE/CVF Conf Comput Vis Pattern Recognition* (2020) 10750–9. doi:10.1109/CVPR42600.2020.01076

14. Ouyang W, Zeng X, Wang X. Modeling mutual visibility relationship in pedestrian detection. *Proc IEEE Conf Comput Vis pattern recognition* (2013) 3222–9. doi:10.1109/CVPR.2013.414

15. Tian Y, Luo P, Wang X, Tang X. Pedestrian detection aided by deep learning semantic tasks. *Proc IEEE Conf Comput Vis pattern recognition* (2015) 5079–87. doi:10.1109/CVPR.2015.7299143

16. Xu D, Ouyang W, Ricci E, Wang X, Sebe N. Learning cross-modal deep representations for robust pedestrian detection. *Proc IEEE Conf Comput Vis pattern recognition* (2017) 5363–71. doi:10.1109/CVPR.2017.451

17. Braun M, Krebs S, Flohr F, Gavrila DM. Eurocity persons: a novel benchmark for person detection in traffic scenes. *IEEE Trans pattern Anal machine intelligence* (2019) 41:1844–61. doi:10.1109/tpami.2019.2897684

18. Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans pattern Anal machine intelligence* (2011) 34:743–61. doi:10.1109/tpami.2011.155

19. Zhang S, Benenson R, Schiele B. Citypersons: a diverse dataset for pedestrian detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017). p. 3213–21.

20. Li G, Zhang S, Yang J. Nighttime pedestrian detection based on feature attention and transformation. In: *2020 25th international conference on pattern recognition (ICPR)*. IEEE (2021). p. 9180–7.

21. Chen YT, Shi J, Mertz C, Kong S, Ramanan D. *Multimodal object detection via bayesian fusion* (2021). *arXiv preprint arXiv:2104.02904*.

22. Jia X, Zhu C, Li M, Tang W, Zhou W. Llvip: a visible-infrared paired dataset for low-light vision. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 3496–504. doi:10.1109/ICCVW54120.2021.00389

23. Li C, Song D, Tong R, Tang M. *Multispectral pedestrian detection via simultaneous detection and segmentation* (2018). *arXiv preprint arXiv:1808.04818*.

24. Liu J, Zhang S, Wang S, Metaxas D. Multispectral deep neural networks for pedestrian detection. (2016) *arXiv preprint arXiv:1611.02644*.

25. Zhang H, Fromont E, Lefèvre S, Avignon B. Guided attentive feature fusion for multispectral pedestrian detection. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (2021). p. 72–80.

26. Zhang L, Liu Z, Zhang S, Yang X, Qiao H, Huang K, et al. Cross-modality interactive attention network for multispectral pedestrian detection. *Inf Fusion* (2019) 50:20–9. doi:10.1016/j.inffus.2018.09.015

27. Zhao B, Wang C, Fu Q. Multi-scale pedestrian detection in infrared images with salient background-awareness. *J Electron Inf Technol* (2020) 42:2524–32. doi:10.11999/JEIT190761

28. Li H, Yang M, Yu Z. Joint image fusion and super-resolution for enhanced visualization via semi-coupled discriminative dictionary learning and advantage embedding. *Neurocomputing* (2021) 422:62–84. doi:10.1016/j.neucom.2020.09.024

29. Xiao W, Zhang Y, Wang H, Li F, Jin H. Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution. *IEEE Trans Instrumentation Meas* (2022) 71:1–15. doi:10.1109/tim.2022.3149101

30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis pattern recognition* (2016) 770–8. doi:10.1109/CVPR.2016.90

31. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. Ssd: single shot multibox detector. In: *Computer vision–ECCV 2016: 14th European conference*, Amsterdam, The Netherlands (Springer) (2016), 21–37.

32. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. *Proc IEEE Conf Comput Vis pattern recognition* (2016) 779–88. doi:10.1109/CVPR.2016.91

33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30.

34. Liu Y, Wang L, Li H, Chen X. Multi-focus image fusion with deep residual learning and focus property detection. *Inf Fusion* (2022) 86-87:1–16. doi:10.1016/j.inffus.2022.06.001

35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. *An image is worth 16x16 words: transformers for image recognition at scale* (2020). *arXiv preprint arXiv:2010.11929.*

36. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers and distillation through attention. *Int Conf machine Learn* (2021) 10347–57.

37. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 10012–22. doi:10.1109/ICCV48922.2021.00986

38. Hu X, Yang K, Fei L, Wang K. Acnet: attention based network to exploit complementary features for rgbd semantic segmentation. In: *2019 IEEE international conference on image processing (ICIP)* (IEEE (2019), 1440–4.

39. Xiang K, Yang K, Wang K. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Opt Express* (2021) 29:4802–20. doi:10.1364/oe.416130

40. Deng F, Feng H, Liang M, Wang H, Yang Y, Gao Y, et al. Feanet: feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In: *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS).* IEEE (2021). p. 4467–73.

41. Zhang J, Liu H, Yang K, Hu X, Liu R, Stiefelhagen R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14679–14694, Dec. 2023. doi:10.1109/TITS.2023.3300537

42. Li H, Yu Z, Mao C. Fractional differential and variational method for image fusion and super-resolution. *Neurocomputing* (2016) 171:138–48. doi:10.1016/j.neucom.2015.06.035

43. Liu Y, Wang L, Cheng J, Li C, Chen X. Multi-focus image fusion: a survey of the state of the art. *Inf Fusion* (2020) 64:71–91. doi:10.1016/j.inffus.2020.06.013

44. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 32–42. doi:10.1109/ICCV48922.2021.00010

45. Zhou D, Liu Z, Wang J, Wang L, Hu T, Ding E, et al. Human-object interaction detection via disentangled transformer. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 19568–77.

46. Xia Z, Pan X, Song S, Li LE, Huang G. Vision transformer with deformable attention. *Proc IEEE/CVF Conf Comput Vis pattern recognition* (2022) 4794–803. doi:10.1109/CVPR52688.2022.00475

47. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. *Training data-efficient image transformers and distillation through attention* (2012). arxiv. 10.48550.

48. Shaw P, Uszkoreit J, Vaswani A. *Self-attention with relative position representations* (2018). *arXiv preprint arXiv:1803.02155.*

49. Li H, Liu J, Zhang Y, Liu Y. A deep learning framework for infrared and visible image fusion without strict registration. *Int J Comput Vis* (2023). doi:10.1007/s11263-023-01948-x

50. Li H, Zhao J, Li J, Yu Z, Lu G. Feature dynamic alignment and refinement for infrared?visible image fusion: translation robust fusion. *Inf Fusion* (2023) 95:26–41. doi:10.1016/j.inffus.2023.02.011

51. Yang Y, Xu K, Wang K. Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection. *Front Phys* (2023) 11: 1–11. doi:10.3389/fphy.2023.1121311

52. Choi Y, Kim N, Hwang S, Kweon IS. Thermal image enhancement using convolutional neural network. In: *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE (2016). p. 223–30.

53. Choi Y, Kim N, Hwang S, Park K, Yoon JS, An K, et al. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans Intell Transportation Syst* (2018) 19:934–48. doi:10.1109/tits.2018.2791533

54. González A, Fang Z, Socarras Y, Serrat J, Vázquez D, Xu J, et al. Pedestrian detection at day/night time with visible and fir cameras: a comparison. *Sensors* (2016) 16:820. doi:10.3390/s16060820

55. Kim N, Choi Y, Hwang S, Kweon IS. Multispectral transfer network: unsupervised depth estimation for all-day vision. *Proc AAAI Conf Artif Intelligence* (2018) 32. doi:10.1609/aaai.v32i1.12297

56. Guan D, Cao Y, Yang J, Cao Y, Tisse CL. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. *Appl Opt* (2018) 57:D108–D116. doi:10.1364/ao.57.00d108

57. Li C, Song D, Tong R, Tang M. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition* (2019) 85:161–71. doi:10.1016/j.patcog.2018.08.005

58. Wagner J, Fischer V, Herman M, Behnke S Multispectral pedestrian detection using deep fusion convolutional neural networks. *ESANN* (2016) 587:509–14.

59. Dollár P, Appel R, Belongie S, Perona P. Fast feature pyramids for object detection. *IEEE Trans pattern Anal machine intelligence* (2014) 36:1532–45. doi:10.1109/tpami.2014.2300479

60. Zhang S, Benenson R, Schiele B Filtered channel features for pedestrian detection. *CVPR* (2015) 1751–1760. doi:10.1109/CVPR.2015.7298784

61. Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection and segmentation. *Proc IEEE Int Conf Comput Vis* (2017) 4950–9. doi:10.1109/ICCV.2017.530

62. Mao J, Xiao T, Jiang Y, Cao Z. What can help pedestrian detection? *Proc IEEE Conf Comput Vis pattern recognition* (2017) 3127–36. doi:10.1109/CVPR.2017.639

63. Wang X, Xiao T, Jiang Y, Shao S, Sun J, Shen C. Repulsion loss: detecting pedestrians in a crowd. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018). p. 7774–83.

64. Zhang S, Wen L, Bian X, Lei Z, Li SZ. Occlusion-aware r-cnn: detecting pedestrians in a crowd. *Proc Eur Conf Comput Vis (Eccv)* (2018) 637–53. doi:10.1007/978-3-030-01219-9_39

65. Zhang L, Zhu X, Chen X, Yang X, Lei Z, Liu Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. *Proc IEEE/CVF Int Conf Comput Vis* (2019) 5127–37. doi:10.1109/ICCV.2019.00523

66. Qingyun F, Dapeng H, Zhaokui W. *Cross-modality fusion transformer for multispectral object detection* (2021). *arXiv preprint arXiv:2111.00273.*

67. Chen YT, Shi J, Ye Z, Mertz C, Ramanan D, Kong S. Multimodal object detection via probabilistic ensembling. *Eur Conf Comput Vis* (2022) 139–58. doi:10.1007/978-3-031-20077-9_9

68. Kim JU, Park S, Ro YM. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Trans Circuits Syst Video Technol* (2021) 32: 1510–23. doi:10.1109/tcsvt.2021.3076466

69. Jocher G, Stoken A, Borovec J, Changyu L, Hogan A, Diaconu L, et al. *ultralytics/yolov5: v3. 0.* Zenodo (2020).

70. Shen Z, Zhang M, Zhao H, Yi S, Li H. Efficient attention: attention with linear complexities. *Proc IEEE/CVF Winter Conf Appl Comput Vis* (2021) 3531–9. doi:10.1109/WACV48630.2021.00357

71. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. Segformer: simple and efficient design for semantic segmentation with transformers. *Adv Neural Inf Process Syst* (2021) 34:12077–90. doi:10.48550/arXiv.2105.15203

72. Li J, Hassani A, Walton S, Shi H. Convmlp: hierarchical convolutional mlps for vision. *Proc IEEE/CVF Conf Comput Vis Pattern Recognition* (2023) 6306–15. doi:10.1109/CVPRW59228.2023.00671

73. Hwang S, Park J, Kim N, Choi Y, So Kweon I. Multispectral pedestrian detection: benchmark dataset and baseline. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015). p. 1037–45.

74. Chen Y, Xie H, Shin H. Multi-layer fusion techniques using a cnn for multispectral pedestrian detection. *IET Comput Vis* (2018) 12:1179–87. doi:10.1049/iet-cvi.2018.5315

75. Guan D, Cao Y, Yang J, Cao Y, Yang MY. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf Fusion* (2019) 50: 148–57. doi:10.1016/j.inffus.2018.11.017

76. Park K, Kim S, Sohn K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition* (2018) 80:143–55. doi:10.1016/j.patcog.2018.03.007

77. Zhuang Y, Pu Z, Hu J, Wang Y. Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection. *IEEE Trans Netw Sci Eng* (2021) 9:1282–95. doi:10.1109/tnse.2021.3139335

78. Zhang Y, Yin Z, Nie L, Huang S. Attention based multi-layer fusion of multispectral images for pedestrian detection. *IEEE Access* (2020) 8:165071–84. doi:10.1109/access.2020.3022623

79. Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In: *Computer vision–ECCV 2020: 16th European conference*. Glasgow, UK (2020). Proceedings, Part XVIII 16 (2020).

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 17 April 2024
DOI 10.3389/fphy.2024.1390236

Application of mixed reality navigation technology in primary brainstem hemorrhage puncture and drainage surgery: a case series and literature review

Xiaoyong Tang[1†], Yanglingxi Wang[1†], Guoqiang Tang[2], Yi Wang[3], Weiming Xiong[1], Yang Liu[1], Yongbing Deng[1*] and Peng Chen[1*]

[1]Department of Neurosurgery, Chongqing Emergency Medical Center, Chongqing University Central Hospital, Chongqing, China, [2]Pre-hospital Emergency Department, Chongqing Emergency Medical Center, Chongqing University Central Hospital, Chongqing, China, [3]Qinying Technology Co., Ltd., Chongqing, China

Check for updates

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Peng Zhang,
First Affiliated Hospital of Wenzhou Medical University, China
Gang Wu,
Fudan University, China
Tao Li,
Nanjing University, China

*CORRESPONDENCE
Yongbing Deng,
✉ 1422495819@qq.com
Peng Chen,
✉ chenpengpph@163.com

†These authors share first authorship

RECEIVED 23 February 2024
ACCEPTED 26 March 2024
PUBLISHED 17 April 2024

CITATION
Tang X, Wang Y, Tang G, Wang Y, Xiong W, Liu Y, Deng Y and Chen P (2024), Application of mixed reality navigation technology in primary brainstem hemorrhage puncture and drainage surgery: a case series and literature review. *Front. Phys.* 12:1390236.
doi: 10.3389/fphy.2024.1390236

Objective: The mortality rate of primary brainstem hemorrhage (PBH) is high, and the optimal treatment of PBH is controversial. We used mixed reality navigation technology (MRNT) to perform brainstem hematoma puncture and drainage surgery in seven patients with PBH. We shared practical experience to verify the feasibility and safety of the technology.

Method: We introduced the surgical procedure of brainstem hematoma puncture and drainage surgery with MRNT. From January 2021 to October 2022, we applied the technology to seven patients. We collected their clinical and radiographic indicators, including demographic indicators, preoperative and postoperative hematoma volume, hematoma evacuation rate, operation time, blood loss, deviation of the drainage tube target, depth of implantable drainage tube, postoperative complications, preoperative and 1-month postoperative GCS, etc.

Result: Among seven patients, with an average age of 56.71 ± 12.63 years, all had underlying diseases of hypertension and exhibited disturbances of consciousness. The average evacuation rate of hematoma was 50.39% ± 7.71%. The average operation time was 82.14 ± 15.74 min, the average deviation of the drainage tube target was 4.58 ± 0.72 mm, and the average depth of the implantable drainage tube was 62.73 ± 0.94 mm. Among all seven patients, four patients underwent external ventricular drainage first. There were no intraoperative deaths, and there was no complication after surgery in seven patients. The 1-month postoperative GCS was improved compared to the preoperative GCS.

Conclusion: It was feasible and safe to perform brainstem hematoma puncture and drainage surgery by MRNT. The technology could evacuate about half of the hematoma and prevent hematoma injury. The advantages included high precision in dual-plane navigation technology, low cost, an immersive operation experience, etc. Furthermore, improving the matching registration method and performing high-quality prospective clinical research was necessary.

KEYWORDS

primary brainstem hemorrhage, mixed reality navigation technology, brainstem hematoma puncture and drainage surgery, neuronavigation, deviation

Introduction

Primary brainstem hemorrhage (PBH) is spontaneous brainstem bleeding associated with hypertension unrelated to cavernous hemangioma, arteriovenous malformation, and other diseases. Hypertension is the leading risk factor for PBH, and other elements include anticoagulant therapy, cerebral amyloid angiopathy, et al. PBH is the deadliest subtype of intracerebral hemorrhage (ICH), accounting for 6%–10% of all ICH with an annual incidence of approximately 2–4/100,000 people [1–3]. The clinical characteristics of PBH are acute onset, rapid deterioration, poor prognosis, and high mortality (30%–90%) [1, 4, 5].

The inclusion criteria of previous ICH research all excluded PBH, such as STICH and MISTIE trials. There is no clear evidence for the optimal treatment of PBH, and the view of surgical treatment has noticeable regional differences. European and North American countries generally believe that severe disability or survival in a vegetative state is a high mental and economic burden for PBH patients and their families. These countries do not favor surgical treatment. However, many PBH surgical treatments have been carried out in China, Japan, and South Korea. Surgical treatment methods, surgical effects, monitoring methods, and complications have been investigated, and much experience has been accumulated.

In 1998, Korean scholars performed the first craniotomy to evacuate the brainstem hematoma [6]. However, in 1989, the Japanese scholar Takahama performed stereotactic brainstem hematoma aspiration surgery [7]. In our opinion, microsurgery craniotomy requires high electrophysiological monitoring and surgical skills, and these limitations are not conductive to popularization. Minimally invasive surgery has the characteristics of a simple operation, minimally invasive, and short operation time, and it is believed to reduce the damage to critical brainstem structures and protect brainstem function as much as possible. More and more minimally invasive treatments have been adopted to improve the precision of PBH puncture, including stereotactic frameworks, robotic-assisted navigation systems, 3D printing techniques, and even laser combined with CT navigation techniques.

Mixed reality navigation technology (MRNT) is based on virtual and augmented reality development. The technology uses CT images to construct a 3D head model and design an individual hematoma puncture trajectory. The actual environmental position is captured by a camera during surgery and was fused with 3D head model synchronously. MRNT not only display the model image combined with actual environment but also navigate the puncture trajectory in real time, allowing the surgeon to precisely control puncture angle and depth to achieve a perfect procedure. This technology makes the head utterly transparent during the surgery and brings an immersive experience to the surgeon.

MRNT has broad application prospects. However, it is still in its infancy, and its application in neurosurgery has rarely been reported. Furthermore, there is no report on application of MRNT in the surgical treatment of PBH. In this study, we used MRNT to perform brainstem hematoma puncture and drainage surgery in seven patients with PBH to share practical experience to verify the feasibility and safety of the technology.

Materials and methods

General information

With the approval of the Ethics Committee of the Chongqing Emergency Medical Center, we included seven patients diagnosed with PBH from January 2021 to October 2022. All underwent brainstem hematoma puncture and drainage surgery with MRNT under general anesthesia. Indications for surgery were patients who 1) were 18–80 years of age; 2) had hematoma volume greater than 5 mL and less than 15 mL; 3) had a diameter of the hematoma greater than 2 cm; 4) had hematoma deviating toward one side or the dorsal side; 5) had GCS less than 8; and 6) had surgery within 6–24 h after onset. Family members were informed and signed the consent form [8]. Exclusion criteria were patients who had 1) brainstem hemorrhage caused by cavernous hemangioma, arteriovenous malformation, and other diseases; 2) GCS >12; 3) bilateral pupil dilation; 4) unstable vital signs; 5) severe underlying disease; or 6) coagulation dysfunction.

Mixed reality navigation technology (MRNT)

All patients preparing for surgery were required to wear sticky analysis markers in the parieto-occipital region and undergo a CT scan before surgery. CT image scanning was performed with a 64-slice CT scanner (Lightspeed VCT 6, General Electric Company, United States of America). The image parameters included in the exposure were 3 mAS, the thickness was 5mm, and the image size was 512 × 512. The DICOM data were analyzed to construct the 3D model of the hematoma and head, and the volume of brainstem preoperative hematoma was calculated using software (Medical Modeling and Design System). In addition, the hematoma puncture trajectory was designed according to the constructed head model.

After general anesthesia, the sticky analysis markers were replaced with bone nail markers, keeping the same position [9]. Based on the principle of near-infrared optical navigation, the camera captured the actual space position in real-time, fused it with the markers of the 3D head model (HSCM3D DICOM), and

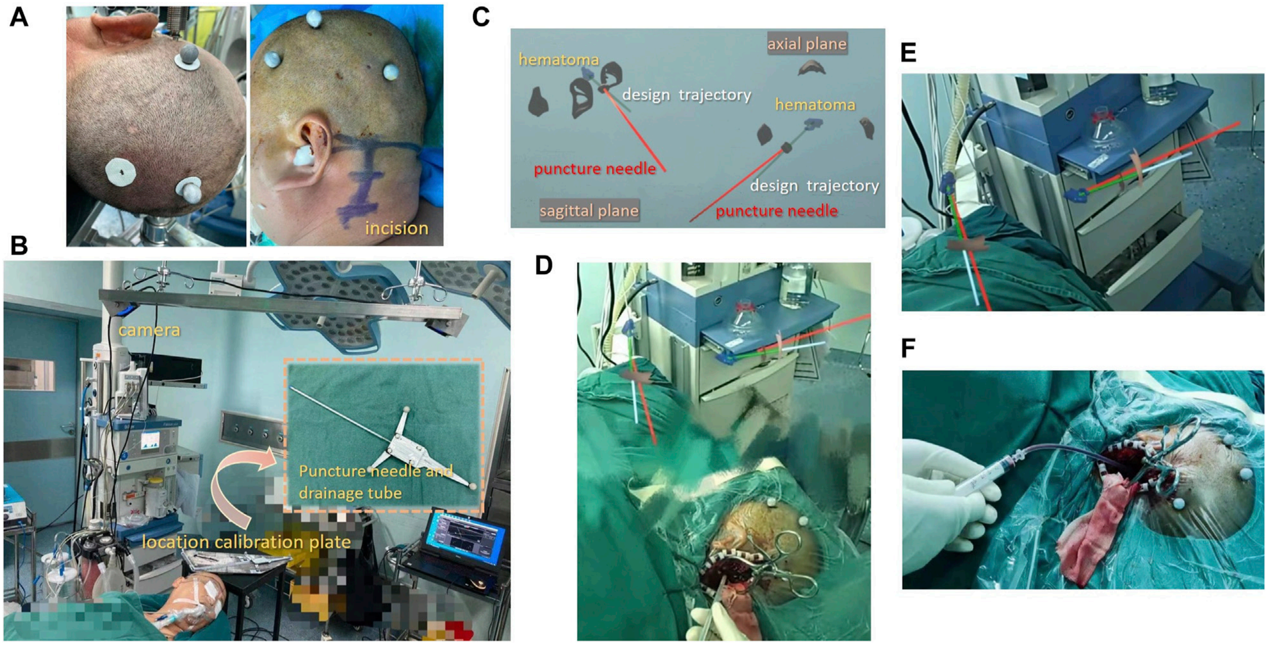


FIGURE 1
Surgical procedure for brainstem hematoma puncture and drainage surgery with MRNT **(A)** Patients were required to wear sticky analysis markers in the parieto-occipital region. **(B)** The camera captured the real space position of the calibration plate, puncture needle, and head. **(C)** Wearing HoloLens, the surgeon viewed the two planes of the image. **(D)** MRNT displays the model image and the actual environment synchronously, allowing the surgeon to perform precise surgery. **(E)** The real-time navigation of MRNT showed that the puncture needle was close to the hematoma target. **(F)** The surgeon was aspirating the hematoma.

transmitted the information to the wearable device (HoloLens). During surgery, the camera continuously tracked the position of the puncture needle to achieve navigation function. In short, the image processing software matched and fused information from camera systems and wearable device through multiple markers. When controlling the movement of surgical tools, the software also processed the dynamic tool position data and fused it with the virtual model through wireless transmission.

Surgical procedures

Hydrocephalus patients were first treated with external ventricular drainage (EVD), and the frontal Kocher point was selected as the cranial entry point. The procedures were cutting the skin, drilling the skull, cutting the dura mater, puncturing in the direction of the plane of binaural connection, fixing the drainage tube, and suturing it layer by layer.

The patient was placed in a prone position with the head frame fixed. The puncture point was 2 cm below the transverse sinus and 3 cm lateral to the midline of the hematoma side. After cutting the skin, the muscle was separated. The dura mater was cut through a drilled hole. Wearing HoloLens, the surgeon synchronously observed actual head structure and fused puncture trajectory from multiple angles and used dual-plane navigation technology [9] for hematoma puncture. After watching that the drainage tube was in place, the puncture needle was removed, and a 5 mL empty syringe was connected for suction. The drainage tube was fixed and sutured layer by layer. The head CT was reviewed immediately after the surgery, and the decision whether to inject urokinase according to the drainage tube's position and the residual hematoma volume. Urokinase was injected from a drainage tube for 2-3 w units every 12 h, usually 4–6 times, and kept for 1.5 h before opening the tube. The retention time of the drainage tube was no more than 72 h after the surgery. The surgical procedure to apply MRNT is shown in Figure 1.

Clinical and radiographic indicators

The indicators for analysis included: demographic indicators, preoperative and postoperative hematoma volume, hematoma evacuation rate, operation time, blood loss, deviation of the drainage tube target, depth of implantable drainage tube, postoperative complications, and preoperative and 1-month postoperative GCS, etc.

The deviation of the drainage tube target was defined as the distance between the tip of the drainage tube and the planned puncture hematoma target. The deviation calculation was done with the BLENDER 2.93.3 software, which used the 3D global coordinate system to visualize the distance.

The head CT examination was reviewed within 24 h after surgery, and the postoperative hematoma volume was measured by non-operators using previous software (Medical Modeling and Design System). Hematoma evacuation rate = (preoperative hematoma volume - postoperative hematoma volume)/preoperative hematoma volume.

TABLE 1 Demographic and clinical characteristics of seven patients.

Case, no	Age (years), gender	Pre-operative volume (ml)	Post-operative volume (ml)	HER (%)	EVD	Deviation (mm)	Depth (mm)	GCS (pre-operation)	GCS (1month)	Outcome
1	67/M	8.54	3.14	63.23	+	4.05	62.57	8	13	CON
2	64/M	5.45	3.18	41.65		4.22	63.42	8	13	CON
3	47/F	6.83	3.51	48.61		4.92	62.92	8	9	COMA
4	37/M	8.36	3.74	55.26	+	5.32	64.23	5	7	COMA
5	53/M	10.21	5.95	41.72	+	3.36	61.84	5	8	COMA
6	55/F	12.21	5.67	53.56	+	4.96	61.42	5		DEAD
7	74/M	7.66	3.93	48.69	+	5.22	62.69	7		GIVE UP

CON, conscious; EVD, external ventricular drainage; GCS, glasgow coma scale; HER, hematoma evacuation rate.

Statistical analysis

All statistical analyses were performed with SPSS (version 21, IBM, Chicago, IL, United States). Quantitative variables are presented as means ± standard deviations. The normality of quantitative variables was assessed through the Kolmogorov-Smirnov test. If the distribution was found to be normal, paired *t*-test were performed. The categorical variables are presented as percentages and tested by χ2 or Fisher's test. A *p*-value less than 0.05 was considered statistically significant.

Results

From January 2021 to October 2022, seven patients were diagnosed with PBH and underwent brainstem hematoma puncture and drainage surgery with MRNT. A summary of the demographic and clinical characteristics of the patients was provided in Table 1. Among the seven patients, five were men, with an average age of 56.71 ± 12.63 years (37–74 years). The seven cases had underlying hypertension, and four cases had diabetes. The average time from onset to admission was 4.2 ± 1.47 h. Seven patients had prominent disturbances of consciousness, four required ventilator assistance, and three had a high fever.

According to the brainstem hematoma classification advocated by Chung [10], 2 cases belonged to small unilateral tegmental type, 4 cases belonged to basal-tegmental type, and other 1 case belonged to bilateral tegmental type. The average volume of preoperative brainstem hematoma was 8.47 ± 2.22 mL (range, 5.45–12.2 mL), the average volume of postoperative brainstem hematoma was 4.16 ± 1.17 mL (range, 3.14–5.95 mL), and the differences were significant. The average hematoma evacuation rate was 50.39% ± 7.71% (range, 41.65%–63.23%). Four of the seven patients underwent EVD first (57.1%), and one underwent EVD 2 days after hematoma puncture and drainage surgery. The average operation time was 82.14 ± 15.74 min, the average blood loss was 32.2 ± 8.14 mL, the average deviation of the drainage tube target was 4.58 ± 0.72 mm (range, 3.36–5.32 mm), and the average depth of the implantable drainage tube was 62.73 ± 0.94 mm (range, 61.42–64.23 mm). Three patients were injected with urokinase after surgery, and the average retention time of the drainage tube was 53.56 ± 7.83 h.

There were no intraoperative deaths in seven patients. Two patients had slight intraoperative fluctuations in vital signs. The most common postoperative comorbidity was pneumonia (7/7, 100%), followed by gastrointestinal bleeding (5/7, 71.43%). There were no rebleeding incidents, ischemic stroke, intracranial infection, or epilepsy within 2 weeks after surgery. The preoperative high fever symptoms were relieved after surgery. Only one patient died due to pneumonia 12 days after surgery, one patient gave up 20 days after surgery. Two patients were conscious and three patients were still in a coma 1 month after surgery.

The average preoperative GCS was 6.57 ± 1.51, and the average postoperative GCS was 10.00 ± 2.83 1 month after surgery. The improvement was statistically significant. The representative cases are shown in Figure 2 and Figure 3.

Discussion

The brainstem is small, deep in the skull, and includes the midbrain, pons, and medulla oblongata. The brainstem is the center of life, controlling respiration, heart rate, blood pressure, and body temperature. About 60%–80% of PBH occurs in the pons due to the rupture of the perforating vessels of the basilar artery [1, 2]. Hypertension is one of the most common causes of severe cerebrovascular disease. By causing mechanical and chemical damage to essential structures in the brainstem, such as the nucleus clusters and the reticular system, the hematoma quickly induces clinical symptoms such as coma, central hyperthermia, tachycardia, abnormal pupils, and hypotension. The prognosis is extremely poor, which presents a challenge to existing treatment methods.

The conservative treatment strategy for PBH is mainly related to the hypertensive treatment strategy for ICH [11]. Since the primary damage of PBH is irreversible, surgical treatment is believed to relieve mechanical compression of the hematoma and prevent secondary injury, improving prognosis [1, 12, 13]. However, there have been some controversies about surgical treatment. Due to the high mortality and disability rate of PBH, it is necessary to strictly evaluate the indications for surgery. Indications for surgery proposed by Shresha included a

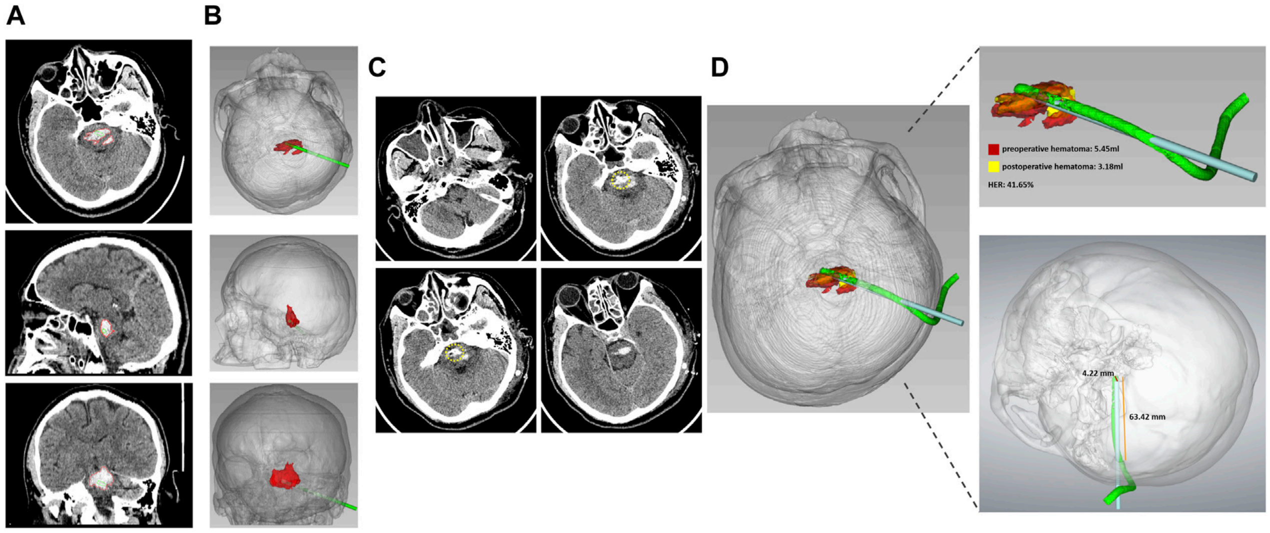


FIGURE 2
The representative case 2 **(A)** Preoperative CT showed PBH in the axial, sagittal, and coronal planes. **(B)** The 3D model constructed from CT images showed hematoma and designed the puncture trajectory from the axial, sagittal, and coronary positions. **(C)** Postoperative CT of the axial plane showed that the drainage tube location was precise. The yellow circle indicated the tip of the drainage tube. **(D)** Fusion of preoperative and postoperative 3D model showed that the preoperative hematoma volume was 5.45 mL, the postoperative hematoma volume was 3.18 mL, the hematoma evacuation rate was 41.65%, the deviation of the target drainage tube was 4.22 mm, and the depth of the implantable drainage tube was 63.42 mm.

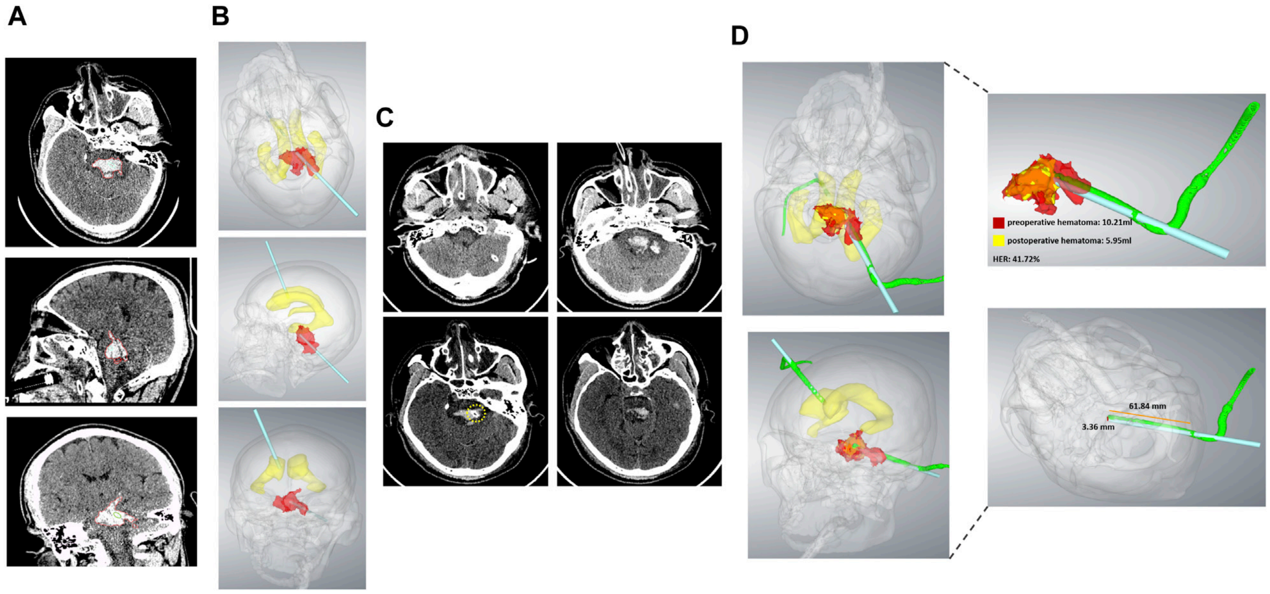

A
B
C
D
preoperative hematoma: 10.21ml
postoperative hematoma: 5.95ml
HER: 41.72%
61.84 mm
3.36 mm

FIGURE 3
The representative case 5. **(A)** Preoperative CT showed PBH in the axial, sagittal, and coronal planes. **(B)** The 3D model constructed from CT images showed hematoma, lateral ventricular, and a designed puncture trajectory from axial, sagittal, and coronary positions. **(C)** Postoperative CT of the axial plane showed that the drainage tube location was precise. The yellow circle indicated the tip of the drainage tube. **(D)** Fusion of the preoperative and postoperative 3D model showed that the preoperative hematoma volume was 10.21 mL, the postoperative hematoma volume was 5.95 mL, the hematoma evacuation rate was 41.72%, the deviation of the drainage tube target was 3.36 mm. The depth of the implantable drainage tube was 61.84 mm.

hematoma volume greater than 5 mL, a relatively concentrated hematoma, GCS less than 8, progressive neurological dysfunction, and uneventful vital signs, particularly requiring ventilatory assistance [14]. Huang established a brainstem hemorrhage scoring system and suggested patients with a score of 2–3 might benefit from surgical treatment. A score of 4 was a contraindication to surgical treatment [15]. A review of 10 cohort studies showed that the patients in the surgical group were 45–65 years old, unconscious, with a GCS of 3–8, and the hematoma volume was approximately 8 mL. The surgical group

TABLE 2 Reported cases of deviations in the application of MR or AR in neurosurgery.

References	Year	Types of studies	Technology	Disease	Deviation (mm)
Chen peng et al. [9]	2022	Case Series	MR	HICH	5.76 ± 0.80
Zhu Tao et al. [28]	2022	Case Series	AR	HICH	1.28 ± 0.43
Zhou Zeyang et al. [27]	2022	Case Series	MR	HICH	Phantom 1.65
					Clinical experiment 1.94
Hou Yuanzheng et al [29]	2016	Case Series	AR	Intracranial Lesions	≤5
Qi Ziyu et al [26]	2021	Case Series	MR	Intracranial Lesions	4.1
Li ye et al [25]	2018	Case Series	MR	EVD	4.34
van Doormaal et al [21]	2019	Case Series	AR	Brain tumor	plastic head 7.2 ± 1.8
					Clinical experiment 4.4 ± 2.5
Tim Fick et al [22]	2021	Meta Analysis	AR	Neurosurgery	4.3

AR, augmented reality; EVD, external ventricular drain; HICH, hypertensive intracerebral hemorrhage; MR, mixed reality.

had a better prognosis and lower mortality than the conservative treatment group. The research also suggested that older age and coma were not contraindications for brainstem hemorrhage surgery [16]. According to the Chinese guidelines for brainstem hemorrhage, we specified the following surgical indications: age 18–80 years old, hematoma volume greater than 5 mL and less than 15 mL, hematoma diameter greater than 2 cm, hematoma deviated to one side or the dorsal side, GCS less than 8, surgery performed within 6–24 h after onset, and family consent [8].

The surgical treatments for PBH included microscopic craniotomy to evacuate the hematoma, which removed the hematoma as much as possible, performed hemostasis, and removed the fourth ventricular hematoma to smooth the circulation of cerebrospinal fluid. However, this technology required various intraoperative monitoring methods and proficient surgical skills. The most widely chosen method was stereotactic hematoma puncture and drainage surgery. To achieve precise puncture of the brainstem hematoma, surgeons had used invasive stereotaxic frames [17], robot-assisted navigation systems [18], the 3D printing technology navigation method [19], and laser combined with CT navigation technology [13]. The above techniques had shortcomings, including invasive placement positioning framework, the risk of skull bleeding and infection, expensive costs of robot-assisted and neuronavigation systems, the lengthy procedure of 3D printing technology, etc.

We innovatively used MRNT to perform brainstem hematoma puncture and drainage surgery. Our team used this technology to successfully perform intracranial foreign body removal [20] and minimally invasive puncture surgery for deep ICH, with a deviation of the drainage tube target of 5.76 ± 0.80 mm [9]. Based on previous experience and technical improvement, we applied technology to perform brainstem hematoma puncture and drainage surgery. The average volume of preoperative brainstem hematoma was 8.47 ± 2.22 mL, postoperative brainstem hematoma was 4.16 ± 1.17 mL, and the average hematoma evacuation rate was 50.39% ± 7.71%, which prevented hematoma primary compression and secondary injury. The surgical procedure under general anesthesia took an average of 82.14 ± 15.74 min, the average target deviation was 4.58 ± 0.72 mm, and the average depth of the implantable drainage tube was 62.73 ± 0.94 mm. The depth of the drainage tube was longer than that in the application of deep ICH, which required higher precision. Moreover, we found MRNT was safe in seven patients.

A comparison of the precision of augmented reality technology, mixed reality technology, and traditional stereotactic methods have been discussed in previous literature. Van Doormaal et al. conducted a holographic navigation study using augmented reality technology. They found that the fiducial registration error was 7.2 mm in a plastic head model, and the fiducial registration error was 4.4 mm in three patients [21]. A meta-analysis was conducted to systematically review the accuracy of augmented reality neuronavigation and compare it with conventional infrared neuronavigation. In 35 studies, the average target registration error of 2.5 mm in augmented reality technology was no different from that of 2.6 mm in traditional infrared navigation [22]. Moreover, In the study of neuronavigation using mixed reality technology, the researchers received a target deviation range of 4–6 mm [23–25].

The augmented reality technology application scenarios mainly involve intracranial tumors and rarely involve ICH. Qi et al. used mixed reality navigation technology to perform ICH surgery. They also used markers for point registration and image fusion. The results showed that the occipital hematoma puncture deviation was 5.3 mm due to the prone and supine position, and the deviation in the basal ganglia was 4.0 mm [26]. Zhou et al. also presented a novel multi-model mixed reality navigation system for hypertensive ICH surgery. The results of the phantom experiments revealed a mean registration error of 1.03 mm. The registration error was 1.94 mm in clinical use, which showed that the system was sufficiently accurate and effective for clinical application [27]. A summary of the deviations in the application of MR or AR was provided in Table 2.

In addition to precision puncture and hematoma drainage, surgical treatment of PBH also required further discussion on the timing of surgery, external ventricular drainage, and fibrinolytic drugs. Shrestha et al. found that surgical treatment within 6 h after onset was associated with a good prognosis [14]. The ultra-early operation alleviated the hematoma mass effect and reduced secondary injury. In particular, for patients with a severe condition, early hematoma aspiration could immediately eliminate harmful effects and prevent worse clinical outcomes

[17] However, many primary hospitals are not equipped with PBH surgical treatment abilities. Patients have to waste a lot of time in the transfer process, which is a big challenge in clinical treatment. PBH can also cause cerebrospinal fluid circulation disorder that induces patients to become unconscious. External ventricular drainage is beneficial in improving cerebrospinal fluid circulation, managing intracranial pressure, and facilitating patient recovery [17]. In our study, external ventricular drainage was performed in five cases of seven patients. Previous research investigating the effects of rtPA on ICH and ventricular hemorrhage by MISTIE and CLAEA demonstrated that fibrinolytic drug administration did not increase the risk of hemorrhage [30–33]. Currently, there is no evidence and consensus to verify the effects of the thrombolytic drug used in PBH. We also found that urokinase did not increase the risk of bleeding and improve drainage efficiency, as reported in previous literature [13, 18].

Compared with the expensive neuronavigation system, mixed reality navigation technology was an independent research and development project, the equipment of the technology was simple, and the cost was low. The effect of the technology met the clinical application of intracerebral hemorrhage surgery, and was beneficial to popularization for primary hospital.

There were also some limitations in our technology. Firstly, in order to introduce our innovative mixed reality navigation technology earlier and faster, we reported few cases, so there are not enough data to verify the advancement of the technology. At present, it was difficult to perform a cohort study because of the small number of patients enrolled. We plan to carry out clinical study with other centers in the future. Secondly, navigation technology was mainly based on point-matching technology, which enabled the fusion of the image model with the actual space through markers. Implementing invasive markers in the skull might carry potential risks of bleeding or infection. Moreover, the procedure required CT examinations before surgery, which delayed surgery time, and increased costs. Some researchers proposed the face registration plan, but the target deviation of the face registration was higher than that of the point registration, and the clinical practicability was poor [34]. Clinical practice must explore a precise, simple, fast, and noninvasive matching and fusion innovative solution.

Conclusion

It was feasible and safe to perform brainstem hematoma puncture and drainage by MRNT. Early minimally invasive precise surgery could prevent hematoma primary and secondary injury, and improve the prognosis of patients with PBH. The advantages included high precision in dual-plane navigation technology, low cost, an immersive operation experience, etc. Furthermore, improving the matching registration method and performing high-quality prospective clinical research was necessary.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by Ethics Committee of the Chongqing Emergency Medical Center. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

XT: Writing–original draft, Data curation, Software. YaW: Writing–original draft. GT: Conceptualization, Project administration, Writing–original draft. YiW: Investigation, Resources, Software, Writing–original draft. WX: Resources, Formal Analysis, Writing–original draft, Writing–review and editing. YL: Methodology, Writing–original draft. YD: Writing–review and editing. PC: Writing–review and editing, Conceptualization, Writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was financially supported by the Fundamental Research Funds for the Central Universities (2022CDJYGRH-015) and Medical Research Project of Science and Technology Bureau and Health Commission, Chongqing, China (2023MSXM076).

Conflict of interest

Author YiW was employed by Qinying Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Chen P, Yao H, Tang X, Wang Y, Zhang Q, Liu Y, et al. Management of primary brainstem hemorrhage: a review of outcome prediction, surgical treatment, and animal model. *Dis Markers* (2022) 2022:1–8. doi:10.1155/2022/4293590

2. Chen D, Tang Y, Nie H, Zhang P, Wang W, Dong Q, et al. Primary brainstem hemorrhage: a review of prognostic factors and surgical management. *Front Neurol* (2021) 12:727962. doi:10.3389/fneur.2021.727962

3. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. *Lancet Neurol* (2010) 9:167–76. doi:10.1016/s1474-4422(09)70340-0

4. Behrouz R. Prognostic factors in pontine haemorrhage: a systematic review. *Eur Stroke J* (2018) 3:101–9. doi:10.1177/2396987317752729

5. Balci K, Asil T, Kerimoglu M, Celik Y, Utku U. Clinical and neuroradiological predictors of mortality in patients with primary pontine hemorrhage. *Clin Neurol Neurosurg* (2005) 108:36–9. doi:10.1016/j.clineuro.2005.02.007

6. Hong JT, Choi SJ, Kye DK, Park CK, Lee SW, Kang JK. Surgical outcome of hypertensive pontine hemorrhages: experience of 13 cases. *J Korean Neurosurg Soc* (1998) 27:59–65.

7. Takahama H, Morii K, Sato M, Sekiguchi K, Sato S. Stereotactic aspiration in hypertensive pontine hemorrhage: comparative study with conservative therapy. *No Shinkei Geka* (1989) 17:733–9.

8. Chen L, Chen T, Mao G, Chen B, Li M, Zhang H, et al. Clinical neurorestorative therapeutic guideline for brainstem hemorrhage (2020 China version). *J Neurorestoratology* (2020) 8:232–40. doi:10.26599/jnr.2020.9040024

9. Peng C, Yang L, Yi W, Yidan L, Yanglingxi W, Qingtao Z, et al. Application of fused reality holographic image and navigation technology in the puncture treatment of hypertensive intracerebral hemorrhage. *Front Neurosci* (2022) 16:850179. doi:10.3389/fnins.2022.850179

10. Chung CS, Park CH. Primary pontine hemorrhage: a new CT classification. *Neurology* (1992) 42(4):830–4. doi:10.1212/wnl.42.4.830

11. Greenberg SM, Ziai WC, Cordonnier C, Dowlatshahi D, Francis B, Goldstein JN, et al. 2022 guideline for the management of patients with spontaneous intracerebral hemorrhage: a guideline from the American heart association/American stroke association. *Stroke* (2022) 53:e282–e361. doi:10.1161/str.0000000000000407

12. Balami JS, Buchan AM. Complications of intracerebral haemorrhage. *Lancet Neurol* (2012) 11:101–18. doi:10.1016/s1474-4422(11)70264-2

13. Wang Q, Guo W, Zhang T, Wang S, Li C, Yuan Z, et al. Laser navigation combined with XperCT technology assisted puncture of brainstem hemorrhage. *Front Neurol* (2022) 13:905477. doi:10.3389/fneur.2022.905477

14. Shrestha BK, Ma L, Lan Z, Li H, You C. Surgical management of spontaneous hypertensive brainstem hemorrhage. *Interdiscip Neurosurg* (2015) 2:145–8. doi:10.1016/j.inat.2015.06.005

15. Huang K, Ji Z, Sun L, Gao X, Lin S, Liu T, et al. Development and validation of a grading Scale for primary pontine hemorrhage. *Stroke* (2017) 48:63–9. doi:10.1161/strokeaha.116.015326

16. Zheng WJ, Shi SW, Gong J. The truths behind the statistics of surgical treatment for hypertensive brainstem hemorrhage in China: a review. *Neurosurg Rev* (2022) 45:1195–204. doi:10.1007/s10143-021-01683-2

17. Du L, Wang JW, Li CH, Gao BL. Effects of stereotactic aspiration on brainstem hemorrhage in a case series. *Front Surg* (2022) 9:945905. doi:10.3389/fsurg.2022.945905

18. Zhang S, Chen T, Han B, Zhu W. A retrospective study of puncture and drainage for primary brainstem hemorrhage with the assistance of a surgical robot. *Neurologist* (2023) 28:73–9. doi:10.1097/nrl.0000000000000445

19. Wang Q, Guo W, Liu Y, Shao W, Li M, Li Z, et al. Application of a 3D-printed navigation mold in puncture drainage for brainstem hemorrhage. *J Surg Res* (2020) 245:99–106. doi:10.1016/j.jss.2019.07.026

20. Li Y, Huang J, Huang T, Tang J, Zhang W, Xu W, et al. Wearable mixed-reality holographic navigation guiding the management of penetrating intracranial injury caused by a nail. *J Digit Imaging* (2021) 34:362–6. doi:10.1007/s10278-021-00436-3

21. van Doormaal TPC, van Doormaal JAM, Mensink T. Clinical accuracy of holographic navigation using point-based registration on augmented-reality glasses. *Oper Neurosurg (Hagerstown)* (2019) 17:588–93. doi:10.1093/ons/opz094

22. Fick T, van Doormaal JAM, Hoving EW, Willems PWA, van Doormaal TPC. Current accuracy of augmented reality neuronavigation systems: systematic review and meta-analysis. *World Neurosurg* (2021) 146:179–88. doi:10.1016/j.wneu.2020.11.029

23. Incekara F, Smits M, Dirven C, Vincent A. Clinical feasibility of a wearable mixed-reality device in neurosurgery. *World Neurosurg* (2018) 118:e422–7. doi:10.1016/j.wneu.2018.06.208

24. McJunkin JL, Jiramongkolchai P, Chung W, Southworth M, Durakovic N, Buchman CA, et al. Development of a mixed reality platform for lateral skull base anatomy. *Otol Neurotol* (2018) 39:e1137–42. doi:10.1097/mao.0000000000001995

25. Li Y, Chen X, Wang N, Zhang W, Li D, Zhang L, et al. A wearable mixed-reality holographic computer for guiding external ventricular drain insertion at the bedside. *J Neurosurg* (2018) 1–8. doi:10.3171/2018.4.JNS18124

26. Qi Z, Li Y, Xu X, Zhang J, Li F, Gan Z, et al. Holographic mixed-reality neuronavigation with a head-mounted device: technical feasibility and clinical application. *Neurosurg Focus* (2021) 51:E22. doi:10.3171/2021.5.focus21175

27. Zhou Z, Yang Z, Jiang S, Zhuo J, Zhu T, Ma S. Surgical navigation system for hypertensive intracerebral hemorrhage based on mixed reality. *J Digit Imaging* (2022) 35:1530–43. doi:10.1007/s10278-022-00676-x

28. Zhu T, Jiang S, Yang Z, Zhou Z, Li Y, Ma S, et al. A neuroendoscopic navigation system based on dual-mode augmented reality for minimally invasive surgical treatment of hypertensive intracerebral hemorrhage. *Comput Biol Med* (2022) 140:105091. doi:10.1016/j.compbiomed.2021.105091

29. Hou Y, Ma L, Zhu R, Chen X, Zhang J. A low-cost iPhone-assisted augmented reality solution for the localization of intracranial lesions. *PLoS One* (2016) 11(7):e0159185. doi:10.1371/journal.pone.0159185

30. Hanley DF, Thompson RE, Rosenblum M, Yenokyan G, Lane K, McBee N, et al. Efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE III): a randomised, controlled, open-label, blinded endpoint phase 3 trial. *Lancet* (2019) 393:1021–32. doi:10.1016/s0140-6736(19)30195-3

31. Hanley DF, Lane K, McBee N, Ziai W, Tuhrim S, Lees KR, et al. Thrombolytic removal of intraventricular haemorrhage in treatment of severe stroke: results of the randomised, multicentre, multiregion, placebo-controlled CLEAR III trial. *Lancet* (2017) 389:603–11. doi:10.1016/s0140-6736(16)32410-2

32. Montes JM, Wong JH, Fayad PB, Awad IA. Stereotactic computed tomographic-guided aspiration and thrombolysis of intracerebral hematoma: protocol and preliminary experience. *Stroke* (2000) 31:834–40. doi:10.1161/01.str.31.4.834

33. Vespa P, McArthur D, Miller C, O'Phelan K, Frazee J, Kidwell C, et al. Frameless stereotactic aspiration and thrombolysis of deep intracerebral hemorrhage is associated with reduction of hemorrhage volume and neurological improvement. *Neurocrit Care* (2005) 2:274–81. doi:10.1385/ncc:2:3:274

34. Mongen MA, Willems PWA. Current accuracy of surface matching compared to adhesive markers in patient-to-image registration. *Acta Neurochir (Wien)* (2019) 161:865–70. doi:10.1007/s00701-019-03867-8

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 26 April 2024
DOI 10.3389/fphy.2024.1391104

Brain functional magnetic resonance imaging in ICU patients who developed delirium

Ren-Jie Song[1†], Fu-Jian Guo[1,2†], Xiao-Fei Huang[1], Mo Li[1], Yi-Yun Sun[1], An-Yong Yu[1]* and Tian-Xi Zhang[1]*

[1]Emergency Department, Affiliated Hospital of Zunyi Medical University, Zunyi, China, [2]Renhuai People's Hospital, Zunyi, China

OPEN ACCESS

EDITED BY
Huafeng Li,
Kunming University of Science and Technology, China

REVIEWED BY
Jun Zhang,
Sixth Medical Center of PLA General Hospital, China
Dongze Li,
Sichuan University, China
Dongfeng Pan,
University of Virginia, United States

*CORRESPONDENCE
An-Yong Yu,
✉ anyongyu@163.com
Tian-Xi Zhang,
✉ 92150@sina.com

†These authors share first authorship

RECEIVED 25 February 2024
ACCEPTED 27 March 2024
PUBLISHED 26 April 2024

CITATION
Song R-J, Guo F-J, Huang X-F, Li M, Sun Y-Y, Yu A-Y and Zhang T-X (2024), Brain functional magnetic resonance imaging in ICU patients who developed delirium.
Front. Phys. 12:1391104.
doi: 10.3389/fphy.2024.1391104

Objective: To detect brain alterations in intensive care unit (ICU) patients who develop delirium using functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI) and to determine their predictive value.

Methods: Fifty-two patients who were admitted to the ICU of the Affiliated Hospital of Zunyi Medical University between June 2021 and June 2022 were enrolled. Fifteen patients who were diagnosed with delirium by the Intensive Care Delirium Screening Checklist (ICDSC) after MRI were selected as the delirium group, and 15 healthy volunteers who were examined during the same period served as the control group. Both groups underwent fMRI and DTI. Quantitative fMRI and DTI data were compared between the two groups to detect abnormal structural and functional brain damage. The relationships between MRI outliers and clinical indicators in the delirium group were also assessed.

Results: Demographic characteristics and imaging indicators before delirium were not correlated with ICDSC scores after delirium. Compared with the healthy control group, the delirium group had significantly lower regional homogeneity (ReHo) values in the left caudate nucleus and frontal lobe on fMRI. The amplitude of the low-frequency fluctuations (ALFF) values of the delirium group were significantly increased in the hippocampus but significantly decreased in the frontal lobe. Compared with the healthy control group, the delirium group showed reduced mean diffusivity (MD) values, mainly in the right cerebellum and right middle temporal gyrus; reduced radial diffusivity (RD) values, mainly in the anterior cerebellum and right middle temporal gyrus; reduced fractional anisotropy (FA) values, only in the corpus callosum; and reduced axial diffusivity (AD) values, mainly in the anterior cerebellar lobe, right middle temporal gyrus, and left middle frontal gyrus on DTI. The statistical thresholds for quantitative DTI measurements were $p < 0.005$ at the voxel level and a cluster size > 5.

Conclusion: Abnormal resting-state brain activity in the left superior frontal gyrus and structural changes in the frontal lobe, temporal lobe, corpus callosum, hippocampus, and cerebellum were observed in ICU patients who developed delirium during hospitalization. Early-brain fMRI and DTI examinations are recommended for the prediction of delirium according to unique quantitative indicators to facilitate early intervention for critically ill patients, reduce the length of hospital stay, and improve patient prognosis.

KEYWORDS

delirium, functional magnetic resonance imaging, diffusion tensor imaging, predictive value, brain imaging

1 Introduction

Delirium is a form of acute organic brain dysfunction that occurs most often in ICU patients. Patients in the ICU are exposed to numerous risk factors for delirium during treatment of the primary illness. Approximately one-third of critically ill patients in the ICU exhibit delirium, and a much greater prevalence of delirium is found in patients receiving mechanical ventilation [1, 2]. Moreover, the incidence of delirium in hospitalized patients aged over 65 years is approximately 48%, and the postoperative incidence is 15%–50% [3]. To date, several studies have shown that delirium is associated with increased mortality, length of hospital stay (LOS), and cost. In addition, when high-risk groups are considered, such as older individuals and patients on mechanical ventilation, delirium can occur in up to 80% of ICU patients [4]. Fluctuations in attention and cognition are observed in patients with delirium [5], and the clinical manifestations, including delusions, hallucinations, or disorientation due to different causes of injury [6], vary and have long-term effects on the quality of life of patients. Despite the high incidence of delirium, it is often underrecognized clinically. Currently, the clinical diagnosis of delirium is usually verified by delirium assessment tools, including the ICU Delirium Assessment Scale, the Intensive Care Delirium Screening Checklist (ICDSC), and the Delirium Screening Checklist [7]. However, these diagnostic tools are highly subjective, time-consuming, and limited by the patient's speech, which poses substantial barriers to the early diagnosis of delirium by clinicians and affects the early treatment and prognosis of patients with delirium.

Previous neuroimaging studies of delirium have shown that patients with cortical atrophy, white matter lesions, and ventricle enlargement are at increased risk of delirium [8]. However, the pathophysiological mechanisms of delirium are still poorly understood. Neuroimaging offers a noninvasive method to advance our understanding of the mechanisms of delirium [9]. Delirium can generally be divided into three types: excitatory, inhibited, and mixed [10]. Excitatory delirium is clinically characterized by excitement and mania, and it is easy to identify clinically. However, inhibited delirium is more common than excitatory delirium. Studies have reported that the proportion of patients with inhibited delirium in the ICU is the highest, followed by patients with mixed delirium and patients with excitatory delirium [11]. Inhibited delirium often leads ICU doctors to misjudge the mental state of patients. Usually, these patients are relatively calm, but patients with severe anxiety and hallucinations are unable to properly express their thoughts. As a result, inhibited delirium is often missed; targeted treatment is often delayed or even absent, and the prognosis of patients with inhibited delirium is often worse.

Delirium has been reported to be correlated with the prognosis of ICU patients, and early intervention is effective in reducing pain [6], disability, mortality, and medical costs. Therefore, there is an urgent need for objective auxiliary tools to assist clinicians in the quick and accurate detection of delirium in patients. Functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI) are newly developed imaging techniques that can measure brain function. These techniques have been widely used to study neurological diseases but have rarely been used to examine individuals with delirium. Previous studies have shown abnormal resting-state activity in several brain regions in patients with delirium. To this end, we conducted resting-state fMRI and DTI examinations of ICU patients in our hospital in this single-center, prospective, cohort study and analyzed patients who developed delirium in the middle and late stages of an ICU admission to further explore the changes in patient-specific brain function in regions before delirium occurred. We hoped to provide a theoretical basis for the use of imaging to facilitate the clinical prediction of delirium, increase early diagnosis and treatment, and reduce the impact of delirium on the quality of life of patients.

2 Data and methods

2.1 Ethics

This study complied with medical ethics standards and was approved by the ethics committee of the Affiliated Hospital of Zunyi Medical University (approval number: KLL-2020-133). All the scanned patients and healthy volunteers were fully informed about the study and volunteered to participate in the present study.

2.2 Study conditions

All imaging scans were performed using the same MRI device (HDxT 3.0 T, GE, USA) at the imaging center of the Affiliated Hospital of Zunyi Medical University under the guidance of specialized technicians.

2.3 Study subjects

From 21 June 2021 to 30 June 2022, 52 patients admitted to the ICU of the Affiliated Hospital of Zunyi Medical University were selected as the study population. All patients underwent fMRI within 72 h after admission to the ICU. Delirium was diagnosed using the ICDSC, which has a total score range of 0–8; scores equal to or greater than 4 indicate clinical delirium. The diagnosis was made by two experienced ICU attending physicians with extensive clinical expertise. All patients included in this study were right-handed based on the Chinese handedness assessment criteria. Healthy volunteers from our hospital during the same period who were matched for age, sex, and educational level with the patients and provided signed informed consent forms indicating their willingness to participate in the study were selected as controls.

2.3.1 Inclusion criteria

The inclusion criteria were as follows: Aged 14–70 years, underwent MRI within 72 h after ICU admission, and had an education level above primary school.

2.3.2 Exclusion criteria

The exclusion criteria were as follows: preexisting neurological and psychiatric disorders or genetic disorders, history of severe hepatorenal or cardiac disorders, presence of metallic medical implants, or deafness or blindness that may affect the outcome of delirium.

2.3.3 Elimination criteria

The elimination criteria were as follows: unwillingness to cooperate with scanning, the use of sedatives, or excessive scan artifacts.

2.4 Data collection

Patient sex, age, education level, and ICDSC score after the occurrence of delirium in the ICU were recorded.

2.5 Image acquisition

A USA (United States) superconducting MRI scanner and skull coil were used. During the scanning process, the subjects were instructed to lie quietly in the supine position, close their eyes, and avoid thinking of anything in particular. Axial T2 fluid-attenuated inversion recovery (FLAIR) images were obtained to rule out organic brain lesions and obvious degradation of white matter. The fMRI parameters were as follows: repetition time (TR) = 2000 m, echo time (TE) = 40 m, slice thickness = 4 mm, slice interval = 0, slice layer = 33, number of slices = 210, field of view (FOV) = 24 cm × 24 cm, number of excitations (NEX) = 1, matrix = 64 × 64, and flip angle = 90°. A total of 6930 images were acquired. The DTI parameters were as follows: TR = 8500 m, TE = 40 m, slice thickness = 4 mm, slice interval = 0, number of slices = 35, FOV = 24 cm × 24 cm, diffusion sensitive gradient (b value) = 0 to 1 000 s/mm^2, diffusion sensitive gradient direction = 25, NEX = 1, matrix = 128 × 128, and flip angle = 90°.

2.6 Data processing

2.6.1 Resting-state fMRI data analysis

Slice timing correction: The time information about each layer of each subject's whole brain (volume) was corrected to eliminate the phase difference of each layer of the time series of each subject. Head motion correction: There was a small amount of head motion caused by the subjects' breathing and heartbeats during data acquisition. Therefore, using the first volume of each subject as the reference standard, the remaining volume was spatially transformed using a six-parameter rigid body transformation to eliminate any head motion. Average images were generated after head motion correction. Spatial normalization: To conduct item-by-item statistical analysis of one or more datasets, voxelwise alignment was performed on all subjects considering the variations in brain shape and size. The average image obtained after head motion correction served as the source image for estimating registration parameters, using the blood oxygen level-dependent (BOLD) brain template in the Montreal Neurological Institute (MNI) space as the reference standard. Subsequently, spatial transformation employing a 12-parameter affine transformation and nonlinear deformation was applied to align the images after head motion correction with normalization to eliminate intersubject differences. Gaussian smoothing: Following spatial normalization, the data were smoothed using an 8-mm full-width at half-maximum Gaussian kernel to further reduce noise. Statistical analysis involved transforming the data to adhere to a normal distribution. Linear drift elimination: Linear drift during data acquisition was removed. Regression: Signals originating from white matter and cerebrospinal fluid were regressed out to mitigate their influence on gray matter signals. Friston's 24-motion parameter model regression was performed to eliminate the influence of head motion on the data. Low-frequency filtering: As the signals related to physiological activity were concentrated in the low-frequency band, the data were bandpass filtered at a frequency range of 0.01–0.1 Hz before statistical analysis. Quantitative calculation: Quantitative indicators such as the amplitude of low-frequency fluctuations (ALFF) and regional homogeneity (ReHo) in all subjects were calculated using DPABI software. Quantitative indicator smoothing: Quantitative indicators were smoothed using a 4-mm full-width at half-maximum Gaussian kernel before statistical analysis. Voxel-by-voxel statistical analysis of quantitative indicators: Based on the generalized linear model, statistical analysis models were constructed to analyze the smoothed quantitative data, and voxel-by-voxel statistical analysis was carried out to establish a model for paired-sample t-tests. Correlation analysis of quantitative indicators: The average value of all quantitative indicators in each brain region was extracted according to the 90 ROIs (including the left and right sides) of the Automated Anatomical Labeling (AAL) brain atlas in the MNI space. Correlation analysis was performed with clinical indicators.

2.6.2 DTI data analysis: data transfer

The DICOM data collected by the device were converted to NIFTI format. Data inspection: The quality of the images was checked individually, mainly to ensure the completeness of the data acquisition and the absence of substantial artifacts. Eddy-current correction and head motion correction: the FMRIB Software Library (FSL) was used to correct for eddy-current effects in the acquired DTI data, as well as for the small amount of head motion caused by the subjects' breathing and heartbeats. Gradient correction: After eddy-current correction, the FSL was used to correct the tensor data. Quantitative calculation: ExploreDTI was applied to calculate quantitative indicators, including mean diffusivity (MD), fractional anisotropy (FA), axial diffusivity (AD), and radial diffusivity (RD). Spatial normalization: To perform item-by-item statistical analysis of one or more sets of data, voxel-by-voxel alignment of all subjects was required, given that the brain shape and size of each subject were different.

With the T2-weighted brain template in the MNI space as the reference standard, the registration parameters were estimated by using the b0 image as the source image after eddy-current correction and head motion correction. Then, spatial transformation of the quantitative indicators of the image was conducted with 12-parameter affine transformation and nonlinear deformation, and the brain images of all subjects were normalized to the template space to eliminate individual differences. Gaussian smoothing: The data were smoothed with an 8-mm full-width at half-maximum Gaussian kernel after spatial normalization to further remove noise. The data were then transformed to follow a normal distribution. Extraction of quantitative indicators: The average values of all quantitative indicators in each brain region were extracted according to the 90 ROIs of the AAL brain atlas in MNI space and the JHU white matter atlas (including the left and right sides) and stored as an MS Excel worksheet. Voxel-by-voxel statistical

TABLE 1 Comparison of demographic characteristics and ICDSC scores between patients who developed delirium and healthy volunteers.

Group	Number	Sex (male: female)	Age	ICDSC score
Delirium group	15	8:7	43.40 ± 10.12	6.27 ± 1.12
Healthy control group	15	8:7	41.5 ± 4.11	1.57 ± 0.63
p-value		>0.05	>0.05	>0.05

Note: ICDSC, Intensive Care Delirium Screening Checklist.

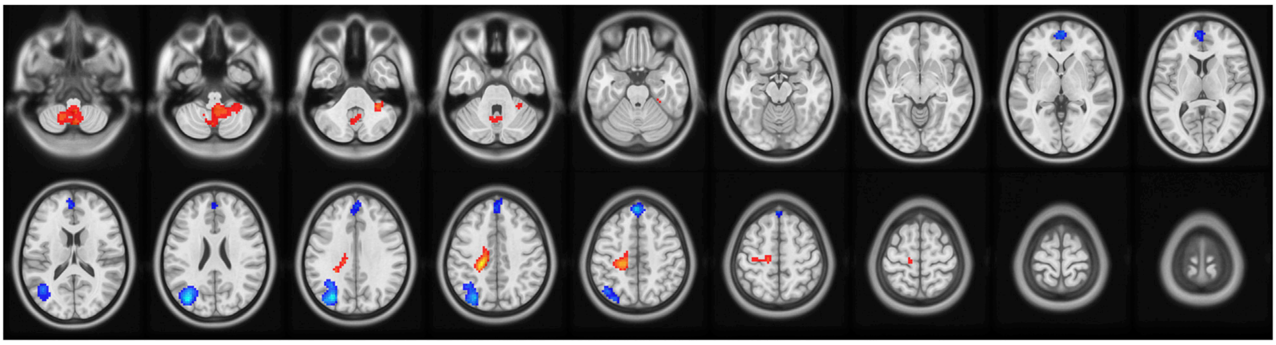


FIGURE 1
Brain regions with abnormal ALFF values in the delirium group. Note: ALFF, amplitude of low-frequency fluctuations; GRF, Gaussian random field. Red indicates brain regions with increased ALFF. Blue indicates brain regions with decreased ALFF.

analysis: Based on the generalized linear model, all quantitative indicators were compared between groups, and the regions with significant differences between groups were identified and stored according to heatmaps and positioning documents. Correlation analysis: A pairwise correlation analysis was performed between all brain regions of all quantitative indicators in the delirium group before and after treatment, and clinical information and p values indicating significant correlations were obtained.

3 Results

3.1 General information and clinical data

Overall, 15 patients were included in the delirium group after scanning. There were eight men and seven women, with an age of 43.40 ± 10.12 years. Fifteen healthy volunteers, including eight men and seven women aged 41.53 ± 4.11 years, composed the control group. There were no statistically significant differences in the numbers of men and women or in the ICDSC scores between the two groups ($p > 0.05$) (Table 1).

3.2 Comparison of ALFF values

ALFF values were analyzed to determine the intensity of spontaneous activity in the voxels identified in the delirium group. Compared with the healthy control group, the delirium group exhibited a significant increase in ALFF values on both sides of the hippocampus (all $p < 0.05$, Gaussian random field [GRF] corrected) and a significant decrease in ALFF values in the frontal lobe (all $p < 0.05$, GRF corrected) (Figure 1; Tables 2, 3).

3.3 Comparison of ReHo values

ReHo values were analyzed to determine the temporal synchronization of local neural activity in the delirium group. Compared with the healthy control group, the delirium group had significantly increased ReHo values in the left brainstem and left medial superior frontal gyrus but significantly decreased ReHo values in the left caudate nucleus and left medial superior frontal gyrus (all $p < 0.05$, GRF corrected) (Figure 2; Tables 4, 5).

3.4 Comparison of MD

MD values were analyzed to determine the changes in regional brain water content and the regional integrity of the myelin sheath in the delirium group. Compared with the healthy control group, the delirium group had decreased MD values in the right cerebellum and right middle temporal gyrus (all $p < 0.05$, GRF corrected) (Figure 3; Table 6).

3.5 Comparison of RD

RD values were analyzed to determine the regional integrity of the myelin sheaths in the delirium group. Compared with the healthy control group, the delirium group had reduced RD values in the anterior cerebellar lobe and right middle temporal gyrus and did not exhibit any regions with increased RD values (Figure 4; Table 7).

TABLE 2 Brain regions with increased ALFF values in patients who developed delirium compared with healthy controls according to resting-state fMRI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Hippocampus	36 37 30	275	<0.05

Note: ALFF, amplitude of low-frequency fluctuation; fMRI, functional magnetic resonance imaging; MNI, Montreal Neurological Institute; GRF Gaussian random field.

TABLE 3 Brain regions with decreased ALFF values in patients who developed delirium compared with healthy controls according to resting-state fMRI.

Brain region	Peak MNI coordinate (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Frontal lobe	−3 54 6	84	<0.05

Note: ALFF, amplitude of low-frequency fluctuations; fMRI, functional magnetic resonance imaging; MNI, Montreal Neurological Institute; GRF Gaussian random field.

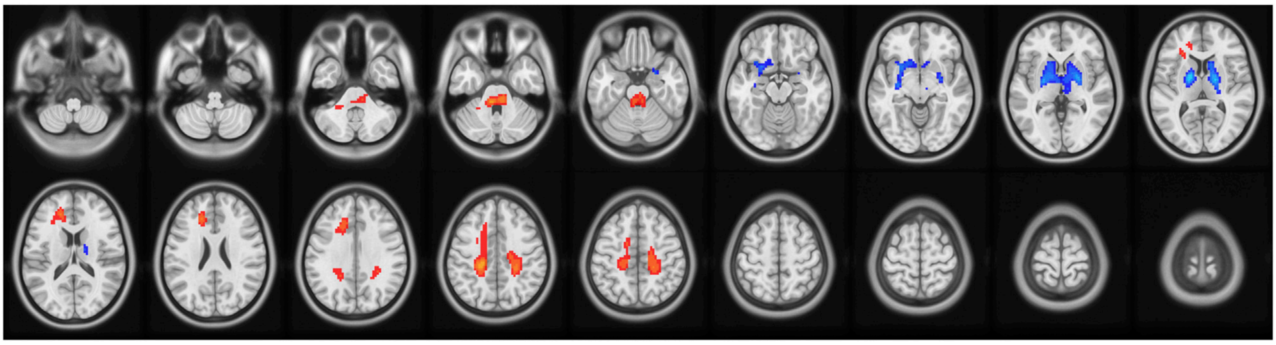


FIGURE 2
Brain regions with abnormal ReHo values in the delirium group. Note: ReHo, regional homogeneity; GRF, Gaussian random field. Red indicates brain regions with increased ReHo. Blue indicates brain regions with decreased ReHo.

TABLE 4 Brain regions with increased ReHo values in patients who developed delirium compared with healthy controls according to resting-state fMRI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Left brainstem	9–36 -30	71	<0.05
Left medial superior frontal gyrus	−20–30 42	23	<0.05

Note: ReHo, regional homogeneity; fMRI, functional magnetic resonance imaging; MNI, Montreal Neurological Institute; GRF, Gaussian random field.

TABLE 5 Brain regions with decreased ReHo values in patients who developed delirium compared with healthy controls according to resting-state fMRI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Left caudate nucleus	−18–3 -6	131	<0.05
Left medial superior frontal gyrus	0 48 48	124	<0.05

Note: ReHo, regional homogeneity; fMRI, functional magnetic resonance imaging; MNI, Montreal Neurological Institute; GRF, Gaussian random field.

3.6 Comparison of FA

Compared with the healthy control group, the delirium group had reduced FA values in the corpus callosum and no areas with no increased FA values (Figure 5; Table 8).

3.7 Comparison of AD

Compared with the healthy control group, the delirium group had reduced AD values in the anterior cerebellar lobe and the left middle frontal gyrus and no regions with no increased AD values (Figure 6; Table 9).

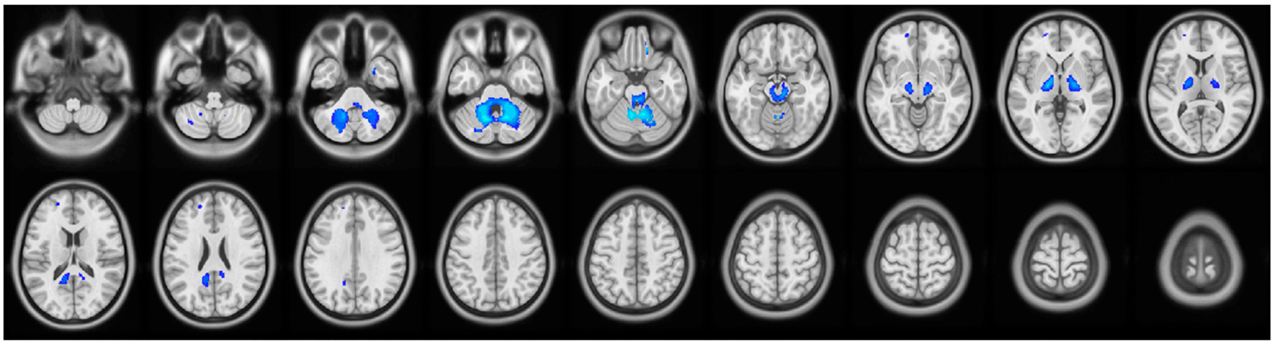


FIGURE 3
Brain regions with abnormal MD values in the delirium group. Note: MD, mean diffusivity. Blue indicates brain regions with decreased MD.

TABLE 6 Brain regions with decreased MD values in patients who developed delirium compared with healthy controls according to DTI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Right cerebellum	−12, −58, −30	374	<0.05
Right middle temporal gyrus	−20, 28, 48	33	<0.05

Note: MD, mean diffusivity; DTI, diffusion tensor imaging; MNI, Montreal Neurological Institute; GRF, Gaussian random field.

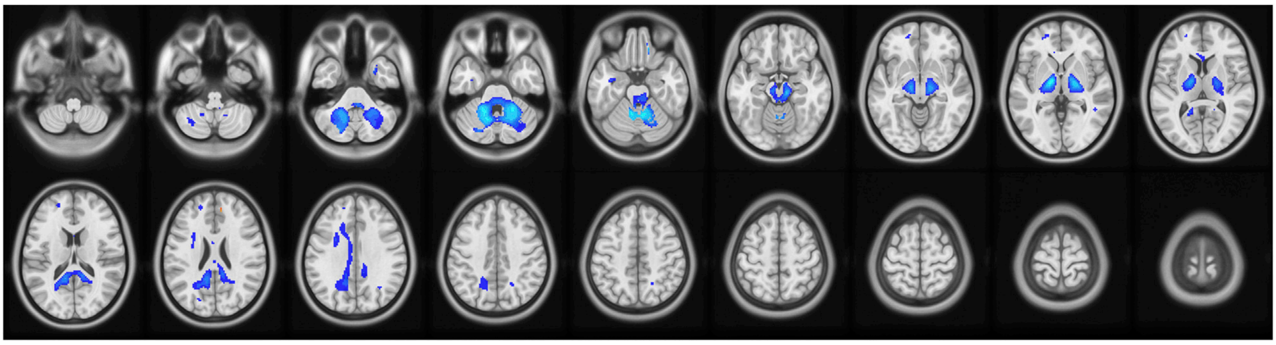


FIGURE 4
Brain regions with abnormal RD values in the delirium group. Note: RD, radial diffusivity. Blue indicates brain regions with decreased RD.

TABLE 7 Brain regions with decreased RD values in patients who developed delirium compared with healthy controls according to DTI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Anterior cerebellar lobe	−14−8 0	2017	<0.05
Right middle temporal gyrus	24 6−40	18	<0.05

Note: RD, radial diffusivity; DTI, diffusion tensor imaging; MNI, Montreal Neurological Institute; GRF, Gaussian random field.

4 Discussion

fMRI can provide a visualization of brain structure and function. Unique quantitative indicators widely used to predict clinical disorders, especially delirium-like disorders, can facilitate early diagnosis and intervention, reduce the length of hospital stay, improve patient prognosis, and decrease mortality. Data, including sex, age, and ICDSC score after the onset of delirium, were collected from patients admitted to the ICU; these variables were not correlated with the imaging indicators before the onset of delirium and thus could not predict whether delirium would occur later. Accordingly, objective variables that can be used to predict

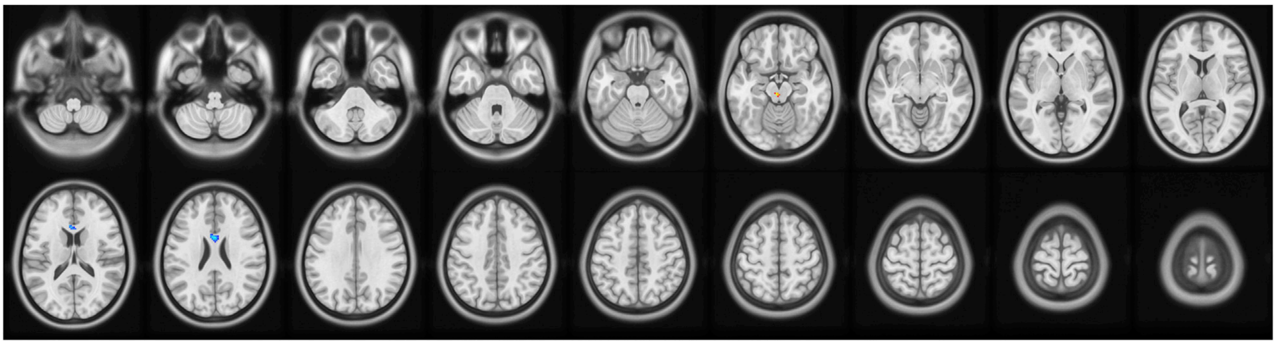


FIGURE 5
Brain regions with abnormal FA values in the delirium group. Note: FA, fractional anisotropy. Blue indicates brain regions with decreased FA.

TABLE 8 Brain regions with decreased FA values in patients who developed delirium compared with healthy controls according to DTI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Corpus callosum	−4–24 14	32	<0.05

Note: FA, fractional anisotropy; DTI, diffusion tensor imaging; MNI, Montreal Neurological Institute; GRF, Gaussian random field.

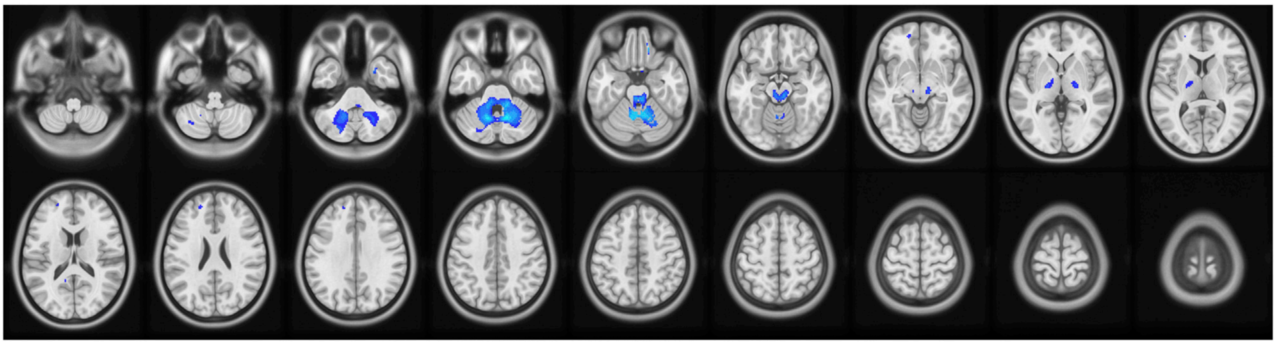


FIGURE 6
Brain regions with abnormal AD values in the delirium group. Note: AD, axial diffusivity. Blue indicates brain regions with decreased AD.

TABLE 9 Brain regions with decreased AD values in patients who developed delirium compared with healthy controls according to DTI.

Brain region	Peak MNI coordinates (mm) (X, Y, Z)	Number of activated voxel clusters	*p*-value, GRF-corrected
Anterior cerebellar lobe	12–56 30	1800	<0.05
Left middle frontal gyrus	−20 48 28	33	<0.05

Note: AD, axial diffusivity; DTI, diffusion tensor imaging; MNI, Montreal Neurological Institute; GRF, Gaussian random field.

delirium are urgently needed. With advances in pathophysiological research on clinical diseases and improvements in DTI and fMRI in recent years, delirium has been studied in depth in the medical field, making it possible to predict the occurrence of delirium.

fMRI analysis methods largely focus on describing the synchrony and spontaneity of brain activity. For instance, ReHo has been utilized to detect similarities in activity between adjacent voxels. ReHo is a whole-brain data analysis method based on Kendall's coefficient of concordance (KCC). The hypothesis proposed by the ReHo method is that brain activity is not shown in a single voxel unit but rather in the form of multivoxel clusters or brain regions. This method uses KCC to measure the synchronization of a particular voxel and its 26 adjacent voxel time series to obtain the KCC map of the whole brain. ALFF can indirectly reflect neuronal activity and describe the intensity of spontaneous voxel activity by calculating the average value of the amplitude at all frequency points within 0.01–0.08 Hz. These two characteristics are the main features of resting-state imaging. In this study, the ALFF and ReHo values in the delirium group were lower than those in the healthy control group prior to the onset of delirium. A previous study showed that the frontal lobe is highly correlated with higher-order functions, such as emotion regulation, cognition, decision making, and executive function [12]. The human prefrontal cortex supports cognitive control, the ability to generate behavioral strategies to coordinate actions and thoughts to achieve internal goals [13]. Lower BOLD signals in the frontal lobe during spontaneous activity and reduced ReHo may impair human emotion regulation, leading to thought disturbances and calculation errors. The results of this study suggest that patients who develop delirium have substantially decreased spontaneous neuronal activity in the frontal lobe compared with healthy volunteers, which is consistent with the findings from the above study. Damage to the left caudate nucleus might be implicated in delirium, as confirmed in our previous study, and thus could lead to the subsequent occurrence of delirium [14]. In this study, the synchronization of neural activity in the left caudate nucleus started to decrease prior to the occurrence of delirium, and activity in brainstem regions was not coordinated, which suggests that this variable may predict the risk of delirium. It was also found that ReHo was increased in the brainstem and that brain connectivity was reduced before the occurrence of delirium, which suggests that these changes may be neural correlates of delirium [15]. Moreover, the ALFF values of the bilateral hippocampus in the delirium group were greater than those in the healthy control group, indicating the involvement of impaired hippocampal function in the occurrence of delirium from an imaging perspective. In addition, animal experiments have demonstrated that the loss of E4bp4 in the hippocampus, which leads to circadian rhythm disturbance, is the basis of the cognitive decline associated with delirium. Pharmacological intervention was shown to affect neuronal activity in the hippocampus and, in turn, cause memory and attention deficits [16]. Taken together, the above findings suggest that impaired hippocampal function is the pathophysiological basis of delirium.

The hippocampus is thought to be strongly associated with consciousness and memory formation [17]. This belief is consistent with our present findings that impaired hippocampal function is involved in the development of delirium. Additionally, we revealed a significant difference in the intensity of spontaneous activity in the hippocampus between the delirium group and the healthy control group, which may provide a foundation for future research on delirium.

The concept of DTI was proposed by Basser et al. in the mid-1990s and has become an important technique in functional magnetic resonance imaging. DTI can further reflect tissue integrity by providing information on the spatial composition of living tissues and water exchange among tissue components under pathological conditions [18]. The MD value is obtained by summing and averaging the three dispersion directions along and perpendicular to the fiber. When a lesion affects brain structure, the integrity of the brain tissue is disrupted, as evidenced by MD, which reflects brain water content and myelin integrity [19]; AD, which reflects axonal integrity; and RD, which reflects myelin integrity. In our study, MD values in the right cerebellum and RD and AD values in the anterior cerebellar lobe were lower in the delirium group before the occurrence of delirium than in the healthy control group, suggesting that the integrity of cerebellar myelination and axons was impaired. This result is consistent with that of another study that identified a strong correlation between dyskinesia and disrupted integrity of cerebellar myelination and axons in patients with delirium [20]. Moreover, neuropsychological and neuroimaging studies have shown that greater cortical function is generally impaired in patients with delirium, particularly in the nondominant prefrontal cortex, frontal cortex, and temporoparietal cortex [21]. Neuronal degeneration and damage to the fiber tract were also observed in the present study; specifically, MD values in the right middle frontal gyrus of the nondominant hemisphere were decreased in the delirium group. FA mainly describes the longitudinal dispersion characteristics of a specific brain region along the direction of the fibers, and a decrease in FA reflects damage to white matter fiber integrity. In our study, FA values in the corpus callosum before the occurrence of delirium were significantly lower in the delirium group than in the healthy control group. Considering that the corpus callosum is the largest commissural fiber network in the brain, disrupted integrity of nerve fibers in the corpus callosum may be the structural basis for subsequent cognitive impairments and personality changes in patients. The corpus callosum is considered the control center for personality abnormalities and cognitive dysfunction; one possible reason is that microstructural changes occur in the upper longitudinal tract following cortical degeneration in patients with delirium, which further leads to various degrees of acute personality abnormalities and cognitive dysfunction [22]. Therefore, the altered FA values of the corpus callosum in the patients who developed delirium observed in our study are consistent with the above studies. Overall, DTI can predict the occurrence of delirium by allowing the direct measurement of the number and integrity of nerve fibers and can reveal the severity of the disease according to patient imaging parameters. Regarding DTI parameters, FA and MD values indicated microstructural damage in multiple brain regions of patients who later developed delirium, and there appeared to be correlations between microstructural damage in different brain regions and the occurrence of delirium. In particular, structural damage to the cerebellum in the same brain region with reduced MD, RD, and AD values seemed to be more closely correlated with the occurrence of delirium.

The main limitation to this study was its relatively small sample size. Further studies with larger sample sizes are required to perform subgroup analyses according to delirium type, such as hyperactive, hypoactive, and mixed types. In addition, the current design of this study does not allow us to establish a true “cause/effect” relationship between delirium and a particular factor.

5 Conclusion

In summary, the abnormal resting-state activity of the left superior frontal gyrus in ICU patients was strongly associated with the subsequent occurrence of delirium. Structural changes and functional abnormalities in the frontal lobe, temporal lobe, corpus callosum, hippocampus, and cerebellum may be preliminary imaging indicators for the prediction of delirium. For patients with delirium, early identification and intervention are recommended to avoid loss of independence and decrease medical costs and mortality risks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by the ethics committee of the Affiliated Hospital of Zunyi Medical University (approval number: KLL-2020-133). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

R-JS: methodology and writing–original draft. F-JG: resources, software, and writing–original draft. X-FH: investigation, methodology, software, and writing–original draft. ML: investigation, methodology, and writing–review and editing. Y-YS: methodology and writing–review and editing. A-YY: funding acquisition, methodology, resources, and writing–review and editing. TZ: conceptualization, data curation, and writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study is part of the "Brain imaging study with functional magnetic resonance imaging in patients with organophosphorus poisoning" project funded by the Science and Technology Support Program of Guizhou Province (registry number: [2021] 047), the National Natural Science Foundation of China (no. 82260385), the Science and Technology Project of Guizhou Province (no. 2021044), the Health Commission of Guizhou Province (gzwkj2022-103), and the Science and Technology Talent Platform of Zunyi [2021] No. 4.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Ely EW. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *JAMA* (2004) 291(14):1753. doi:10.1001/jama.291.14.1753

2. Girard TD, Exline MC, Carson SS, Hough CL, Rock P, Gong MN, et al. Haloperidol and ziprasidone for treatment of delirium in critical illness[J/OL]. *New Engl J Med* (2018) 379(26):2506–16. doi:10.1056/NEJMoa1808217

3. Salluh JI, Soares M, Teles JM, Ceraso D, Raimondi N, Nava VS, et al. Delirium epidemiology in critical care (DECCA): an international study. *Crit Care* (2010) 14(6): R210. doi:10.1186/cc9333

4. Devlin JW, Fong JJ, Howard EP, Skrobik Y, McCoy N, Yasuda C, et al. Assessment of delirium in the intensive care unit: nursing practices and perceptions. *Am J Crit Care* (2008) 17:555–65. quiz 566. doi:10.4037/ajcc2008.17.6.555

5. American Psychiatric Association (APA). *Washington (2020) Diagnostic and statistical manual of mental disorders*. 5th ed. APA.

6. Prendergast NT, Tiberio PJ, Girard TD. Treatment of delirium during critical illness[J/OL]. *Annu Rev Med* (2022) 73(1):407–21. doi:10.1146/annurev-med-042220-013015

7. Devlin JW, Marquis F, Riker RR, Robbins T, Garpestad E, Fong JJ, et al. Combined didactic and scenario-based education improves the ability of intensive care unit staff to recognize delirium at the bedside. *Crit Care* (2008) 12(1):R19. doi:10.1186/cc6793

8. Soiza RL, Sharma V, Ferguson K, Shenkin SD, Seymour DG, Maclullich AM. Neuroimaging studies of delirium: a systematic review[J/OL]. *J Psychosomatic Res* (2008) 65(3):239–48. doi:10.1016/j.jpsychores.2008.05.021

9. Nitchingham A, Kumar V, Shenkin S, Ferguson KJ, Caplan GA. A systematic review of neuroimaging in delirium: predictors, correlates and consequences[J/OL]. *Int J Geriatr Psychiatry* (2018) 33(11):1458–78. doi:10.1002/gps.4724

10. Grover S, Sharma A, Aggarwal M, Mattoo SK, Chakrabarti S, Malhotra S, et al. Comparison of symptoms of delirium across various motoric subtypes[J]. *Psychiatry Clin neurosciences*. 2014,68(4):283–91. doi:10.1111/pcn.12131

11. Bellelli G, Morandi A, Di Santo SG, Mazzone A, Cherubini A, Mossello E, et al. Delirium Day": a nationwide point prevalence study of delirium in older hospitalized patients using an easy standardized diagnostic tool[J]. *BMC Med*. 2016,14:106. doi:10.1186/s12916-016-0649-8

12. Shamay-Tsoory SG, Aharon-Peretz J. Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study[J/OL]. *Neuropsychologia* (2007) 45(13):3054–67. doi:10.1016/j.neuropsychologia.2007.05.021

13. Duverne S, Koechlin E. Rewards and cognitive control in the human prefrontal cortex[J/OL]. *Cereb Cortex* (2017) 27(10):5024–39. doi:10.1093/cercor/bhx210

14. Song R, Song G, Xie P, Duan H, Zhang T, Lu Y, et al. Diffusion tensor imaging and resting-state functional magnetic resonance imaging in patients with delirium in intensive care unit. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue* (2020) 32(1):88–93. doi:10.3760/cma.j.cn121430-20190905-00016

15. Boukrina O, Kowalczyk M, Koush Y, Kong Y, Barrett AM. Brain network dysfunction in poststroke delirium and spatial neglect: an fMRI study[J/OL]. *Stroke* (2022) 53(3):930–8. doi:10.1161/STROKEAHA.121.035733

16. Ferrier J, Tiran E, Deffieux T, Tanter M, Lenkei Z. Functional imaging evidence for task-induced deactivation and disconnection of a major default mode network hub in the mouse brain[J/OL]. *Proc Natl Acad Sci* (2020) 117(26):15270–80. doi:10.1073/pnas.1920475117

17. Wang Q, Zhang X, Guo YJ, Pang YY, Li JJ, Zhao YL, et al. Scopolamine causes delirium-like brain network dysfunction and reversible cognitive impairment without neuronal loss[J/OL]. *Zoolog Res* (2023) 44(4):712–24. doi:10.24272/j.issn.2095-8137.2022.473

18. Huang J, Friedland RP, Auchus AP. Diffusion tensor imaging of normal-appearing white matter in mild cognitive impairment and early alzheimer disease: preliminary evidence of axonal degeneration in the temporal lobe[J/OL]. *Am J Neuroradiology* (2007) 28(10):1943–8. doi:10.3174/ajnr.A0700

19. Westman E, Cavallin L, Muehlboeck JS, Zhang Y, Mecocci P, Vellas B, et al. Sensitivity and specificity of medial temporal lobe visual ratings and multivariate regional MRI classification in alzheimer's disease. *PLoS ONE* (2011) 6(7):e22506. doi:10.1371/journal.pone.0022506

20. Machado AS, Darmohray DM, Fayad J, Marques HG, Carey MR. A quantitative framework for whole-body coordination reveals specific deficits in freely walking ataxic mice. *eLife* (2015) 4:e07892. doi:10.7554/eLife.07892

21. Burns A. Delirium[J/OL]. *J Neurol Neurosurg Psychiatry* (2004) 75(3):362–7. doi:10.1136/jnnp.2003.023366

22. Yuan JL, Wang SK, Guo XJ, Teng LL, Jiang H, Gu H, et al. Disconnections of cortico-subcortical pathways related to cognitive impairment in patients with leukoaraiosis: a preliminary diffusion tensor imaging study[J/OL]. *Eur Neurol* (2017) 78(1–2):41–7. doi:10.1159/000477899

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 10 May 2024
DOI 10.3389/fphy.2024.1397732

OPEN ACCESS

EDITED BY
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China

REVIEWED BY
Tao Ye,
China University of Mining and Technology, China
Wangxia Zuo,
University of South China, China
Guanqiu Qi,
Arizona State University, United States

*CORRESPONDENCE
Jiangwei Li,
✉ jiangweili2023@163.com

RECEIVED 08 March 2024
ACCEPTED 19 April 2024
PUBLISHED 10 May 2024

CITATION
Wang G, Li J, Tan H and Li X (2024), Fusion of full-field optical angiography images via gradient feature detection.
Front. Phys. 12:1397732.
doi: 10.3389/fphy.2024.1397732

Fusion of full-field optical angiography images via gradient feature detection

Gao Wang[1], Jiangwei Li[2]*, Haishu Tan[2] and Xiaosong Li[2]

[1]State Key Laboratory of Dynamic Measurement Technology, North University of China, Taiyuan, China, [2]School of Physics and Optoelectronic Engineering, Foshan University, Foshan, China

Full-field optical angiography (FFOA)—a real-time non-invasive imaging technique for extracting biological blood microcirculation information—contributes to an in-depth understanding of the functional and pathological changes of biological tissues. However, owing to the limitation of the depth-of-field (DOF) of optical lenses, existing FFOA imaging methods cannot capture an image containing every blood-flow information. To address this problem, this study develops a long-DOF full-field optical angiography imaging system and proposes a novel multi-focus image fusion scheme to expand the DOF. First, FFOA images with different focal lengths are acquired by the absorption intensity fluctuation modulation effect. Second, an image fusion scheme based on gradient feature detection in a nonsubsampled contourlet transform domain is developed to capture focus features from FFOA images and synthesize an all-focused image. Specifically, FFOA images are decomposed by NSCT into coefficients and low-frequency difference images; thereafter, two gradient feature detection-based fusion rules are used to select the pre-fused coefficients. The experimental results of both phantom and animal cases show that the proposed fusion method can effectively extend the DOF and address practical FFOA image defocusing problems. The fused FFOA image can provide a more comprehensive description of blood information than a single FFOA image.

KEYWORDS

full-field optical angiography, gradient feature detection, multi-focus image fusion, nonsubsampled contourlet transform, fusion rule

1 Introduction

Blood microcirculation information is critical for gaining insights into both the normal development and pathogenesis of diseases such as cancer and diabetic retinopathy [1–3]; for example, microvascular rarefaction is a hallmark of essential hypertension [4]. Therefore, it is essential to accurately depict high-resolution full-field images of blood vessels to enhance the accuracy of biological studies. In existing full-field optical imaging methods, such as full-field optical coherence tomography [5], laser scatter contrast imaging [6], and full-field optical angiography (FFOA) [7, 8], the imaging speed and sensitivity of bio-optical imaging can be slightly improved, but the imaging range is limited to the depth-of-field (DOF). In addition, high-resolution images are usually obtained by increasing the magnification of the lens, which further reduces the DOF range and cannot ensure that all relevant objects in focus are distinctly imaged. The multi-focus image fusion technique is a feasible method for addressing the issue of a limited DOF. Images of the same scene with different DOFs can be obtained by changing the focal length; thereafter, the focus features from these images are extracted to synthesize a sharp image to extend the DOF.

Current multi-focus image fusion methods can be essentially classified into four categories [9]: transform domain [10–13], spatial domain [14–19], sparse representation (SR) methods [20–25], and deep learning methods [26–30]. The spatial domain methods implement image fusion mainly by detecting the activity level of pixels or regions. For example, Xiao et al. [31] used the multi-scale Hessian matrix to acquire the decision maps. SAMF [32]proposes a new small-area-aware algorithm for enhancing object detection capability. MCDFD [33]proposes a new scheme based on multi-scale cross-differencing and focus detection for blurred edges and over-sharpening of fused images. Spatial domain methods are known for their simplicity and speed; however, accurately detecting pixel activity poses a significant challenge. Inaccurate pixel activity detection may lead to block artifact occurrence and introduce spectral distortions of the fusion results. Since the overcomplete dictionaries of SR methods contain richer basis atoms, SR methods are more robust to misalignment than spatial domain methods [34]. Tang et al. [35] used joint patch grouping and informative sampling to build an overcomplete dictionary for SR. SR is usually time-consuming, and sparse coding using SR is complex; furthermore, it inevitably loses important information of source images. Recently, deep learning methods have gained widespread attention owing to their excellent feature representation capabilities. Liu et al. [26] first applied a CNN to obtain the initial decision of focused and out-of-focus regions. Thereafter, other authors proposed extensive deep learning image fusion algorithms, including generative adversarial network-based [36], encoder-decoder-network based [37], and transform-based methods [27]. REOM [38] measure the similarity between the source images and the fused image based on the semantic features at multiple abstraction levels by CNN. AttentionFGAN [39] used dual discriminators in order to avoid the modal unevenness caused by a single discriminator. Tang et al. [40] proposed an image fusion method based on multiscale adaptive transformer, which introduces adaptive convolution to perform convolution operation to extract global contextual information. CDDFuse [41] propose a novel correlation-driven feature decomposition fusion network, to tackle the challenge in modeling cross-modality features and decomposing desirable modality-specific and modality-shared features. However, these training data lack consistency with real multi-focal images; therefore, real multi-focal images cannot be processed effectively. Transform domain methods decompose images into different scales, analogous to the process of human eyes handling visual information ranging from coarse to fine; thus, the latter can achieve a better signal-to-noise ratio [42]. Transform domain methods usually include pyramid transform [43], wavelet transform [44, 45], and nonsubsampled contourlet transform (NSCT) [46, 47].

In a previous study, a large-DOF FFOA method was developed that uses the contrast pyramid fusion algorithm (CPFA) to achieve image fusion [48]. Pyramid transform is a popular tool that is simple and easy to implement; however, it creates redundant data in different layers and easily loses high-frequency details. In comparison with the pyramid transform, the wavelet transform has attracted more attention owing to its localization, direction, and multi-scale properties. Nevertheless, discrete wavelet transform cannot accurately represent anisotropic singular features [16]. Because it is flexible, multi-scale, multi-directional, and sift-invariant, NSCT has gained an encouraging reputation for multi-focus image fusion and can decompose images in multiple directions and obtain fusion results with more correct information. Li et al. [16] performed comprehensive experiments to analyze the performance of different multi-scale transforms in image fusion and their experimental results demonstrated that the NSCT can overperform other multi-scale transforms in terms of multi-focus image fusion. This study devised a long-DOF full-field optical technique based on gradient feature detection (GFD). A series of FFOA images with different focal lengths were first acquired by the absorption intensity fluctuation modulation (AIFM) effect [8]. Subsequently, a novel multi-focus image fusion method in the NSCT domain was developed to fuse the source FFOA images to extend the DOF. The proposed fusion scheme includes the following three steps. First, the initial images (FFOA images with different DOFs) are decomposed by NSCT into corresponding low-frequency coefficients (LFCs); thereafter, a series of high-frequency directional coefficients (HFDCs), and low-frequency difference images (LFDIs) are obtained by subtracting the LFCs from the source images. Second, two gradient feature detection-based fusion rules are proposed to select the pre-fused coefficients. Finally, the fused image is generated by taking the inverse NSCT (INSCT) on different pre-fused coefficients. This article compared the fusion results using objective assessment and subjective visual evaluation. The experimental results show that the proposed GFD fusion scheme can yield better blood microcirculation images and effectively retain the focus information in the source image.

The main contributions of this study are as follows:

(1) This article constructs a full-field optical imaging system to acquire phantom and animal FFOA images with different DOFs.
(2) This article proposes a gradient feature detection-based image fusion scheme in the NSCT domain that can effectively fuse FFOA images to extend the DOF.
(3) This article develops two fusion rules to fuse the LFCs and HFDCs of NSCT that can be used to extract more detailed and structured FFOA image information, thereby improving the visual perception of the fused images.

The remainder of this paper is organized as follows. Section 2 introduces the imaging system, acquisition of FFOA images, and proposed fusion model based on GFD in the NSCT domain. Section 3 focuses on the experimental results and discussion. Finally, Section 4 provides the conclusions of the study.

2 Materials and methods

2.1 System setup

A schematic of the constructed system is given in Figure 1. The 80-mW laser beam (λ0 = 642 nm, bandwidth = 10 nm) from the semiconductor is reflected by the beam splitter (BS), thus vertically illuminating the sample; the speckle pattern is recorded by a

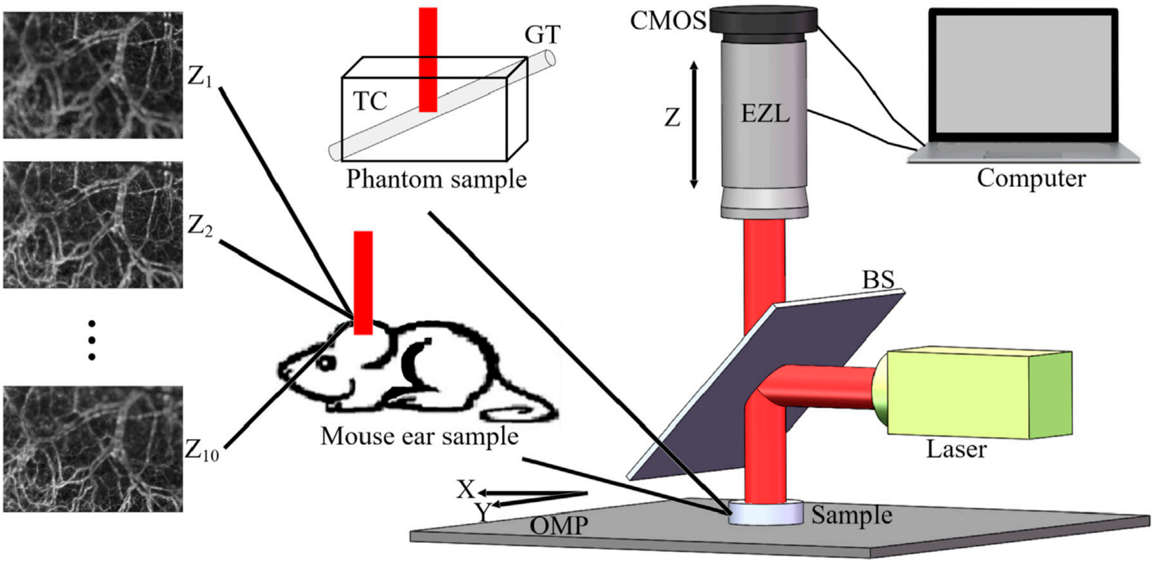


FIGURE 1
GFD FFOA fusion system. Z1 to Z10 represent 10 FFOA images with different foci, OMP is the optical mobile platform, BS is the beam splitter, EZL is electric zoom lens, TC is the transparent container, and GT is the glass tube.

complementary metal-oxide semiconductor (CMOS) camera (acA2000-340km, Basler. Pixel size, 5.5 μm × 5.5 μm; sampling rate, 42 fps; exposure time, 20 ms). Samples were placed on the optical mobile platform (OMP), and the focal length in the *z*-direction was changed by computer control of the electric zoom lens (EZL) to obtain FFOA images (Z_1, Z_2, Z_3, $\ldots$ Z_{10}) with different DOFs; the multi-focus image fusion technique was then used to fuse the 10 images to obtain the fused image. The data collection was controlled using LabVIEW software.

2.2 Acquisition of FFOA image

First, describe the theory of the AIFM effect in realizing the FFOA image [8]. Under irradiation from a low-coherence light source, the red blood cell (RBC) absorption coefficient is significantly higher than the background tissue. In the vascular region, when the RBCs flow, a high-frequency fluctuation signal (IAC) is generated by the combination of different absorptions of RBCs and background tissue; the above phenomenon is called the AIFM effect. However, the region outside the blood vessels produces a DC signal (IDC) that does not fluctuate over time because it only contains background tissue. Thereafter, the time sequences (IAC) and (IDC) are independently demodulated by respectively applying a high-pass filter (HPF) and low-pass filter (LPF) in the frequency domain. The employed formulas are as Eq. (1):

$$\begin{array}{c} I_{DC}(x,y,t) = LPF\{I(x,y,t)\} \\ I_{AC}(x,y,t) = HPF\{I(x,y,t)\} \end{array} \tag{1}$$

where $I(x,y,t)$ is the value of the pixel at spatial coordinate (x,y) at time *t*. The samples have a small concentration of scattering examples, so the collected intensity signal is proportional to the scattering concentration, i.e., $I_{DC} \propto n_{DC}$ and $I_{AC} \propto n_{AC}$, where n_{AC} and n_{DC} represent the moving RBC and background scattering numbers, respectively. Under the condition of $I_{AC} \ll n_{AC}$, the moving RBC concentration can be defined as Eq. (2):

$$\rho = \frac{n_{AC}}{n_{AC}+n_{DC}} \approx \frac{n_{AC}}{n_{DC}} = \frac{I_{AC}}{I_{DC}} \tag{2}$$

In current FFOA methods [7, 8], the imaging parameter is called averaged modulation depth (*AMD*), defined as the ratio of the average dynamic signal intensity $\langle I_{AC}(x,y,t)\rangle_t$ to the average static signal intensity $\langle I_{DC}(x,y,t)\rangle_t$. The employed formula is as Eq. (3):

$$AMD(x,y) = \frac{\langle I_{AC}(x,y,t)\rangle_t}{\langle I_{DC}(x,y,t)\rangle_t} \tag{3}$$

2.3 Proposed fusion scheme for FFOA images

The proposed fusion scheme is illustrated in Figure 2. For a convenient explanation, only two FFOA images are used for the entire process, and the above process is iterated to achieve the fusion of three or even more images. The fusion scheme mainly includes three steps. First, the NSCT is performed on the source images to obtain the corresponding LFCs and HFDCs, and LFDIs are obtained by subtracting the LFCs from the source images. Thereafter, a sum-modified-Laplacian and local energy (*SMLE*) is used to fuse the LFCs, and the structural tensor and local sharpness change metric (*SOLS*) is used to process the LFDIs to obtain the initial decision map. Finally, the HFDC of the fused image is obtained by fusing the HFDC obtained by the final decision map, and an INSCT is performed on all coefficients to generate the final fused image.

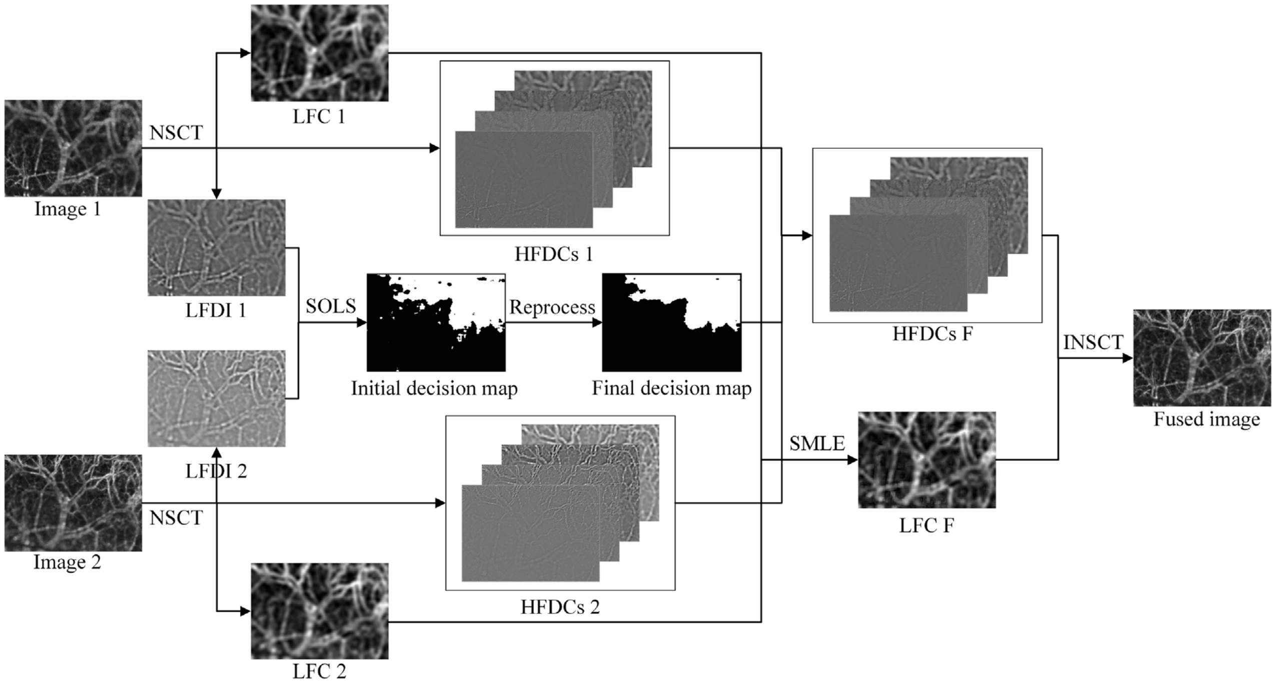


FIGURE 2
Proposed FFOA images fusion scheme.

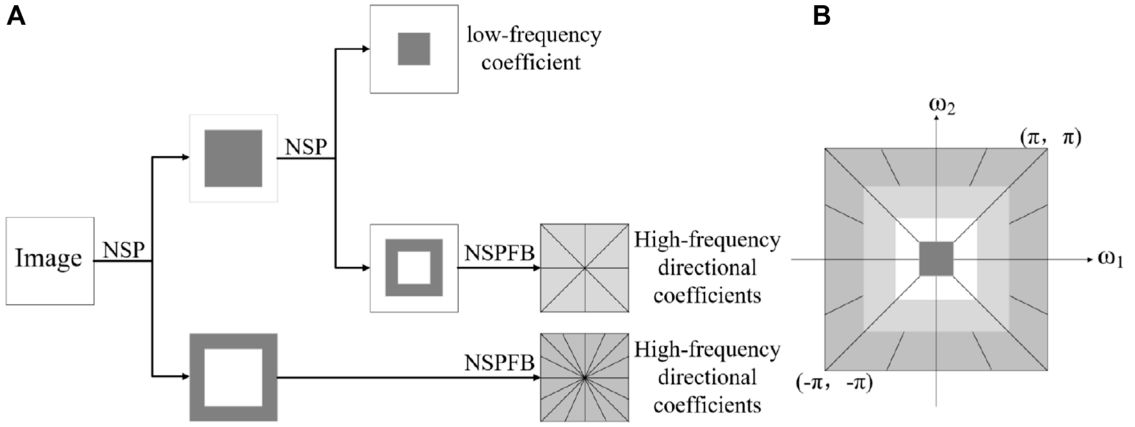

A
B
low-frequency coefficient
NSP
Image
NSP
NSPFB
High-frequency directional coefficients
NSPFB
High-frequency directional coefficients
ω2
(π, π)
ω1
(-π, -π)

FIGURE 3
Overview of NSCT **(A)** Nonsubsampled filter bank structure. **(B)** Idealized frequency partitioning.

2.3.1 NSCT

The NSCT consists of a non-subsampled pyramid (NSP) structure and non-subsampled directional filter banks (NSDFBs) to provide a decomposition of images [47]. Figure 3 depicts an overview of the NSCT. The ideal support regions of the low-frequency and high-frequency filters at the j level are complementary and can be expressed as $[-(\pi/2^j), (\pi/2^j)]^2$ and $[-(\pi/2^{j-1}), (\pi/2^{j-1})]^2/[-(\pi/2^j), (\pi/2^j)]^2$, respectively. The source image is first decomposed into a high-frequency coefficient (HFC) and an LFC by NSP; subsequently, the LFC is decomposed iteratively using NSP. After processing by k-stage NSP, k+1 coefficients (an LFC and k HFCs) with the same size as the source image are generated.

The k-th level NSP is defined as Eq. (4):

$$H_n(Z) = \begin{cases} H_1\left(Z^{2^{n-1}I}\right)\prod_{j=0}^{n-2} H_0\left(2^{2jI}\right), & 1 \le n \le k \\ \prod_{j=0}^{n-2} H_0\left(2^{2jI}\right), & n = k+1 \end{cases} \tag{4}$$

where $H_n(Z)$ is the low-pass filter, and $H_n(Z)$ is the high-pass filter at the n-th stage. NSDFB is a filter bank consisting of a two-channel tree structure. The HFCs from the NSP are decomposed by the NSDFB in one step, and one HFC can generate 2l HFDCs. Because upsampling and downsampling are eliminated, NSDFB can provide directional unfolding with shift invariance for the image. Further details about the NSCT can be found in [47].

To understand the following presentation, let us recall some frequently used symbols. A, B, and X denote the source images, F indicates the fused image, and (x, y) represents the pixel points in the image. The LFC and HFC of the source image X are represented by $C_L^F(x, y)$ and $C_{g,l}^F(x, y)$, respectively, where L represents the coarsest scale, and g and l are the decomposition level and direction, respectively. $\overline{I_X(x, y)}$ denote the low frequency difference image and is obtained by subtracting the LFC from the original image $(\overline{I_X(x, y)} = X(x, y) - C_L^X(x, y))$.

2.3.2 LFCs fusion based on SMLE

In addition to the selection of transform domain, fusion rules are also critical in the multi-focus fusion method. For a pair of LFCs of image X obtained by NSCT decomposition, which retains the majority of the energy information from the source image, the energy change between the clear and defocused objects in the image is relatively large. According to existing literature [49], sum-modified-Laplacian (SML) performs excellently in guiding the selection of LFCs. SML is defined as Eq. (5):

$$SML\left(C_L^X(x, y)\right) = \sum_{m=-M}^{M}\sum_{n=-N}^{N} ML_L^X(x+m, y+n)^2 \tag{5}$$

where $M \times N$ denotes the 3×3 window centered at (x, y). $ML_L^X(x, y)$ denotes the modified Laplacian of $C_L^X(x, y)$ at point (x, y), and is defined as Eq. (6):

$$\begin{aligned} ML_L^X(x, y) = &\left|2C_L^X(x, y) - C_L^X(x-1, y) - C_L^X(x+1, y)\right| \\ &+ \left|2C_L^X(x, y) - C_L^X(x, y-1) - C_L^X(x, y+1)\right| \end{aligned} \tag{6}$$

SML can effectively reflect the changes in the energy of LFCs but cannot reflect the brightness information; therefore, adding the local energy (LE) of LFCs is considered to improve SML. The LE is defined as Eq. (7):

$$LE\left(C_L^X(x, y)\right) = \sum_{m=-M}^{M}\sum_{n=-N}^{N} C_L^X(x+m, y+n)^2 \tag{7}$$

where $M \times N$ denotes the window size centered at (x, y); considering the time complexity and performance, they can be set as $M = N = 1$. Therefore, a combination of SML and LE is used to construct a new fusion rule (*SMLE*), as shown in Eq. (8):

$$SMLE\left(C_L^X(x, y)\right) = SML\left(C_L^X(x, y)\right) \star LE\left(C_L^X(x, y)\right) \tag{8}$$

SMLE is selected as the fusion rule, and the coefficient with a larger *SMLE* is taken as the LFC after fusion. The coefficient selection principle for an LFC can be described as Eq. (9):

$$C_L^F(x, y) = \begin{cases} C_L^A(x, y), & SMLE\left(C_L^A(x, y)\right) > SMLE\left(C_L^B(x, y)\right) \\ C_L^B(x, y), & otherwise \end{cases} \tag{9}$$

where $C_L^F(x, y)$ denotes the LFC of the fused image, $C_L^A(x, y)$ and $C_L^B(x, y)$ are the LFCs of images A and B decomposed by NSCT, respectively.

2.3.3 HFDCs fusion based on SOLS

The process of HFDC fusion based on *SOLS* consists of three steps. First, the initial decision map is obtained by describing the changes in LFCs using *SOLS*; thereafter, the initial decision map is optimized using consistency verification and morphological filtering operations to obtain the final decision map; and finally, the final decision map is used to guide the fusion of HFDCs.

The HFDCs obtained by the NSCT decomposition mainly contain most of the detailed information, such as contours, lines, edges, region boundaries, and textures, and the local geometric structures (LGS) of the focused region tend to be more prominent [50]. Therefore, fusion can be achieved by describing the variation of LGS in HFDC. In recent years, the SOT has gained widespread adoption in image fusion, emerging as a critical method for analyzing the LGS of images [51]. This article selected SOT as a descriptive tool to describe the variation of LGS in the HFDC; however, when SOT is directly selected to guide HFDC fusion, the decision maps of different HFDCs may not be consistent, which can lead to the introduction of error information in the fused images; thus, the fusion decision maps of HFDCs of decision maps are obtained by LFDIs. The process steps are described as follows.

Considering the low frequency difference image $\overline{I_X(x, y)}$ of image X, the square of the rate of change of image A in any direction θ at the point (x, y) can be expressed as [52]:

$$\begin{aligned} \left(d\overline{I_X}\right)^2 &= \left\|\overline{I_X}(x + \varepsilon\cos\theta, y + \varepsilon\sin\theta) - \overline{I_X(x, y)}\right\|_2^2 \\ &\approx \sum_{\omega(x,y)} \left(\frac{\partial\overline{I_X}}{\partial x}\varepsilon\cos\theta + \frac{\partial\overline{I_X}}{\partial y}\varepsilon\sin\theta\right)^2 \end{aligned} \tag{10}$$

where the window $\omega(x, y)$ is defined as the Gaussian function $exp - \frac{(x^2+y^2)}{2\delta^2}$. Using $C(\theta)$ to represent the change rate of LGS of image $\overline{I_X(x, y)}$, Eq. (10) can be expressed as Eq. (11):

$$\begin{aligned} C(\theta) &= \sum_{\omega(x,y)} \left(\frac{\partial\overline{I_X}}{\partial x}\varepsilon\cos\theta + \frac{\partial\overline{I_X}}{\partial y}\varepsilon\sin\theta\right)^2 \\ &= (\cos\theta, \sin\theta)\begin{bmatrix} \sum\limits_{\omega(x,y)}\left(\frac{\partial\overline{I_X}}{\partial x}\right)^2 & \sum\limits_{\omega(x,y)}\frac{\partial\overline{I_X}}{\partial x}\frac{\partial\overline{I_X}}{\partial y} \\ \sum\limits_{\omega(x,y)}\frac{\partial\overline{I_X}}{\partial x}\frac{\partial\overline{I_X}}{\partial y} & \sum\limits_{\omega(x,y)}\left(\frac{\partial\overline{I_X}}{\partial y}\right)^2 \end{bmatrix}(\cos\theta, \sin\theta)^T \\ &= (\cos\theta, \sin\theta)\sum_{\omega(x,y)}\nabla g\nabla g^T(\cos\theta, \sin\theta)^T \end{aligned} \tag{11}$$

where $\nabla g = (\frac{\partial\overline{I_X}}{\partial x}\ \frac{\partial\overline{I_X}}{\partial y})^T$; $\nabla g\nabla g^T$ is the *SOT*, which is defined as Eq. (12):

$$S = \sum_{\omega(x,y)}\nabla g\nabla g^T = \begin{bmatrix} H & M \\ M & V \end{bmatrix} \tag{12}$$

where $H = \sum_{\omega(x,y)}(\frac{\partial\overline{I_X}}{\partial x})^2$, $M = \sum_{\omega(x,y)}\frac{\partial\overline{I_X}}{\partial x}\frac{\partial\overline{I_X}}{\partial y}$, and $V = \sum_{\omega(x,y)}(\frac{\partial\overline{I_X}}{\partial y})^2$. The structure tensor S has two eigenvalues, which can be explicitly calculated as Eq. (13):

$$\lambda_{1,2} = \frac{1}{2}\left((H+V) \pm \sqrt{(V-H)^2 + 4M^2}\right) \tag{13}$$

In general, relatively small values of λ_1 and λ_2 indicate that pixel values change minimally in the region, i.e., they are flat. A larger value of λ_1 or λ_2 indicates a large change in the pixel in one direction, and this region is more inclined to be the focusing region. The structure tensor [53] salient detection operator can be defined as Eq. (14):

$$STS\left(\overline{I_X(x,y)}\right) = \sqrt{(\lambda_1+\lambda_2)^2 + 0.5(\lambda_1-\lambda_2)^2} \tag{14}$$

STO can describe the amount of LGS information in LFDIs; however, it cannot accurately reflect the changes in local contrast. In this study, the sharpness change metric (*SCM*) is used to overcome this deficiency, and the *SCM* is defined as Eq. (15):

$$SCM\left(\overline{I_X(x,y)}\right) = \sum_{(x_0,y_0\in\Omega_0)} \left(\overline{I_X(x,y)} - \overline{I_X(x_0,y_0)}\right)^2 \tag{15}$$

In the formula, Ω_0 is a local region of size 3 × 3 centered on (x,y). In addition, considering the correlation between the pixels in the (x,y) neighborhood, the local *SCM* (*LSCM*) is optimized as Eq. (16):

$$LSCM\left(\overline{I_X(x,y)}\right) = \sum_{m=-M}^{M}\sum_{n=-N}^{N} SCM(x+m,y+n) \tag{16}$$

where $M \times N$ is a neighborhood with size of 3×3.Therefore, a combination of *SOT* and *LSCM* is used to construct a new fusion rule (*SOLS*), as shown in Eq. (17):

$$SOLS\left(\overline{I_X(x,y)}\right) = STS\left(\overline{I_X(x,y)}\right) \times LSCM\left(\overline{I_X(x,y)}\right) \tag{17}$$

Consequently, the process of constructing the initial decision map $IDM(x,y)$ of the fused image detail layer using *SOLS* can be described as Eq. (18):

$$IDM(x,y) = \begin{cases} 1, & SOLS\left(\overline{C_A^L(x,y)}\right) > SOLS\left(\overline{C_B^L(x,y)}\right) \\ 0, & otherwise \end{cases} \tag{18}$$

where $SOLS\left(\overline{C_A^L(x,y)}\right)$ and $SOLS\left(\overline{C_B^L(x,y)}\right)$ denote the *SOLS* of the LFDIs A and B, respectively. The *IDM* in Figure 2 reveal small holes, fine grooves, protrusions, and narrow cracks. To correct these erroneous pixels, the "*bwareaopen*" filter with adaptive threshold was utilized to improve the *IDM*, as described in Eq. (19):

$$MDM = bwareaopen(IDM(x,y), th) \tag{19}$$

where MDM denotes the intermediate decision map. The "*bwareaopen*" filter removes isolated areas smaller than the threshold (*th*) in the binary map. Considering that different image sizes adapt to different values of *th*, *th* = 0.015 × S in our scheme, where S denotes the image area. Considering the object integrity, the MDM can be further improved using the consistency verification operations., as described in Eq. (20):

$$FDM(x,y) = \begin{cases} 1, \ if \displaystyle\sum_{(a,b)\in\Theta} MDM(x+a,y+b) \\ 0, \qquad otherwise \end{cases} \tag{20}$$

where $FDM(x,y)$ is the final decision map of the detail layer, and Θ is a square neighborhood centered at (x,y) with size 21 × 21.

The fusion detail layer is generated using the final decision map as Eq. (21):

$$C_{g,l}^F(x,y) = \begin{cases} C_{g,l}^A(x,y), \ if & FDM(x,y) = 1 \\ C_{g,l}^B(x,y), & otherwise \end{cases} \tag{21}$$

where $C_{g,l}^F(x,y)$ denotes the HFDC of the fused image, and $C_{g,l}^A(x,y)$ and $C_{g,l}^B(x,y)$ are HFDCs of images A and B decomposed by NSCT, respectively.

Finally, the fused image is obtained by INSCT using the LFC $C_L^F(x,y)$ and HFDCs $C_{g,l}^F(x,y)$.

2.4 Evaluation of the FFOA images

For subjective visual evaluation, this article measured the quality of fusion using the difference image, which was obtained by subtracting the fused image from the source image; the difference image $D_n(x,y)$ is given as Eq. (22):

$$D_{\mathrm{n}}(x,y) = F(x,y) - I_{\mathrm{n}}(x,y) \tag{22}$$

where $F(x,y)$ denotes the final fused image, and $I_{\mathrm{n}}(x,y)$ denotes the n-th source image. This article inverted the pixel value of the information residual image for better observation.

For the same focused regions in the fused images, less residual information in the difference image indicates better performance of the fusion method; therefore, difference images are employed for subjective visual evaluation.

Subjective visual evaluation offers a direct comparison, but occasionally, it may be difficult to determine the best performing case. In contrast, objective evaluations can provide a quantitative analysis of fusion quality. In this study, six popular metrics were used to evaluate fusion quality: 1) Normalized Mutual Information (Q_{MI}) [54] for measuring the information preservation degree; 2) Nonlinear Correlation Information entropy (Q_{NCIE}) [55] for measuring the nonlinear correlation degree; 3) Gradient-based Fusion Performance (Q_G) [56]; 4) Image Fusion Metric-Based on a Multiscale Scheme (Q_M) [57] for measuring image features; 5) Metric-Based on Phase Congruency (Q_P) [58]; and 6) Visual Information Fidelity (VIF) [59]. Considering the evaluation results of these metrics, a comprehensive evaluation of the fusion effect can be provided. The greater the value of all these metrics, the better the quality of the fused image. Further information regarding the calculation of objective evaluations can be found in [60].

The proposed method was compared with four advanced methods—CPFA [48], IFCNN [61], U2Fusion [62], and NSSR [63]— to verify its effectiveness. For a fair comparison, the parameter settings of all the methods were consistent with the original publications. In the fusion experiments, the CPFA, NSSR and proposed methods were implemented in MATLAB 2019a, IFCNN and U2Fusion methods were implemented in PyCharm 2022. All the fusion methods were executed on a PC using an Intel(R) Core (TM) i7-5500U CPU @ 2.40 GHz (2,394 MHz) and 12 GB RAM.

3 Results and discussion

To verify the effectiveness of the GFD scheme, this article compared the CPFA, IFCNN, U2Fusion, NSSR, and GFD using

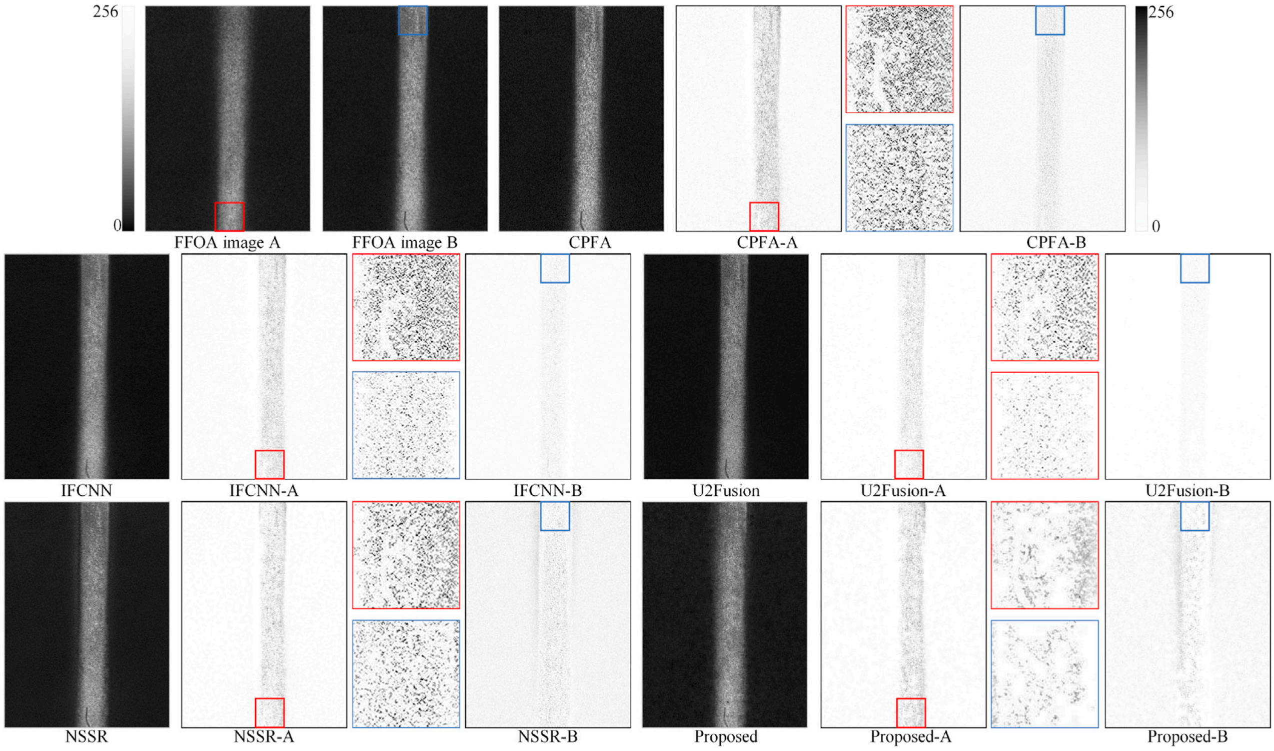


FIGURE 4
Subjective evaluation of phantom experiments.

phantom and animal experimental results. In all examples, using 10 images with different DOFs for fusion, the DOF was extended by multiples of three. For the NSCT used in the proposed method, a four-layer decomposition was performed from coarse to fine in [1, 1, 1, 1] directions, with "vk" and "pyrexc" as the pyramid and the direction filter, respectively. In addition, owing to the limited DOF extended by the fusion of the two images, in all experiments, this article chose to fuse 10 images to get the final fusion results.

3.1 Phantom experimental

This article first demonstrates the validity of the method through simulation experiments, the experimental results of which are shown in Figure 4. A glass tube with a 0.15 mm radius was placed inside the transparent container at an angle of 60° to the horizontal for simulating blood vessels, and the transparent container was filled with 3.2 mg/mL of agar solution to imitate background tissue. RBCs were simulated using an approximately 5 μm radius TiO_2 particle, and 0.5-mg/mL TiO_2 solution was injected into the glass tube at a speed of 5 mL/h to simulate blood flow. In the experiment, the EZL increased the focal length by 2.4 mm each time to acquire FFOA images; the magnification of the lens was 2, the camera exposure time and frame rate were 0.8 millisecond and 95 fps, respectively, and the DOF was expanded from 1 to ~3.2 mm.

Figure 4 shows the FFOA fusion results generated by the different methods. FFOA images A and B represent the first and 10th images, with the focus regions in the images boxed in red and blue, respectively. The fusion results of each method contain three images; the first image represents the fused image produced by fusing 10 FFOA images, and the second and third images are the difference images produced by subtracting the fused image from FFOA images A or B, respectively. The magnified image of the boxed region was placed in the middle of the two difference images for better visibility. By analyzing the red and blue boxed regions, it can be found that the proposed method and U2Fusion had fewer residuals; in contrast, the difference images of CPFA, IFCNN, and NSSR had more residual information. The above results indicate that the proposed method can retain more source image information than other methods. Figure 5 was obtained by excluding the subjective visual evaluation in Figure 4; it was used to validate the effectiveness of the proposed method and shows the objective evaluation metrics of the nine fusions used in the phantom experiment. Furthermore, it shows that in the objective evaluation of Q_{MI}, Q_G, Q_M, and Q_P, both NSSR and the proposed method exhibited excellent performance; however, the proposed method was slightly better than the NSSR method, whereas the CPFA, IFCNN, and U2Fusion performed poorly in these objective evaluations. Regarding the objective evaluation VIF, NSSR performed the best, and the proposed method and U2Fusion also showed good performance. The subjective and objective evaluations in Figures 4, 5 showed that the proposed method is effective in the phantom experiment.

3.2 Animal experimental

This article performed vivo experiments using mouse ears to validate the proposed method further. The mouse (C57BL/6,

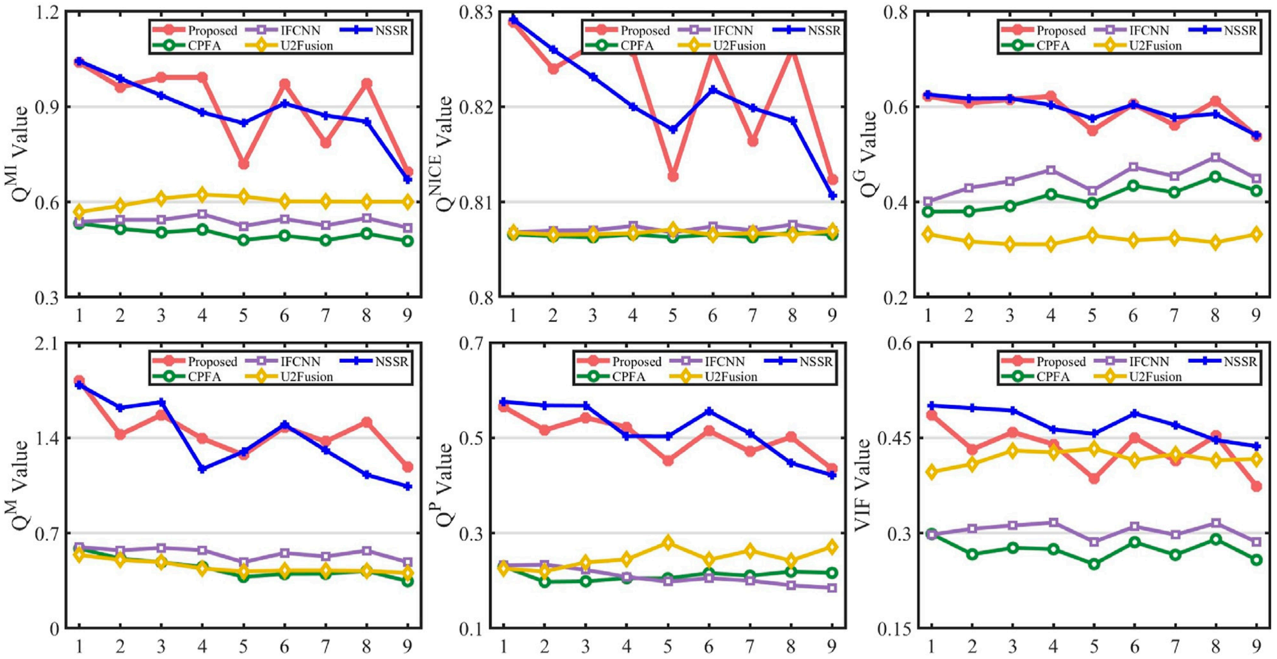


FIGURE 5
Objective evaluation of phantom experiments.

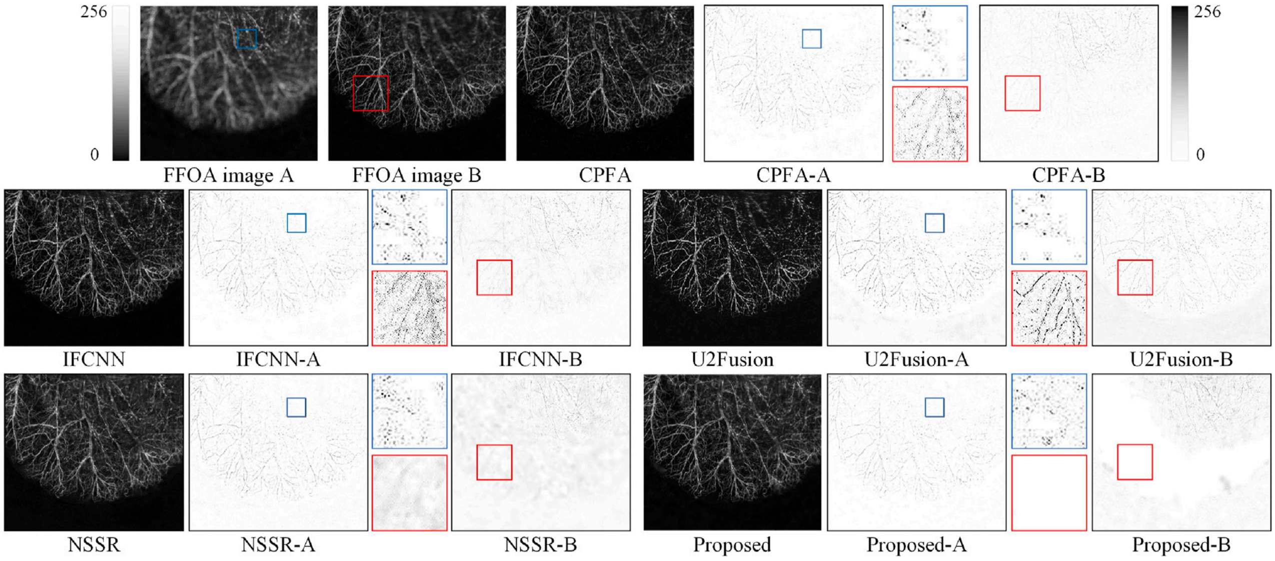


FIGURE 6
Subjective evaluation of the first of group mouse ear experiments.

9 weeks old, and 21 g in weight) was anesthetized with 0.12 mL of chloral hydrate at a concentration of 0.15 g/mL. In the experiment, the EZL increased the focal length by 2.4 mm each time to acquire FFOA images, the magnification of the lens was 1.15, and the camera exposure time and frame rate were 0.45 ms and 42 fps, respectively. For a fair comparison, the experimental data of the first group mouse ear is from literature [48], and the second group mouse ear is from literature [63].

Figures 6, 7 show the experimental results of two different groups of mouse ears. The DOF was expanded from 0.8 to ~3.3 mm. Figure 6 presents the first group of mouse ear experiments, and the FFOA images A and B are mouse ears with different DOFs; the focused regions are marked with red and blue boxes. This article boxed some blood vessels with different thicknesses in FFOA image A and one complete vascular vein in FFOA image B. The fusion results of each method contain three

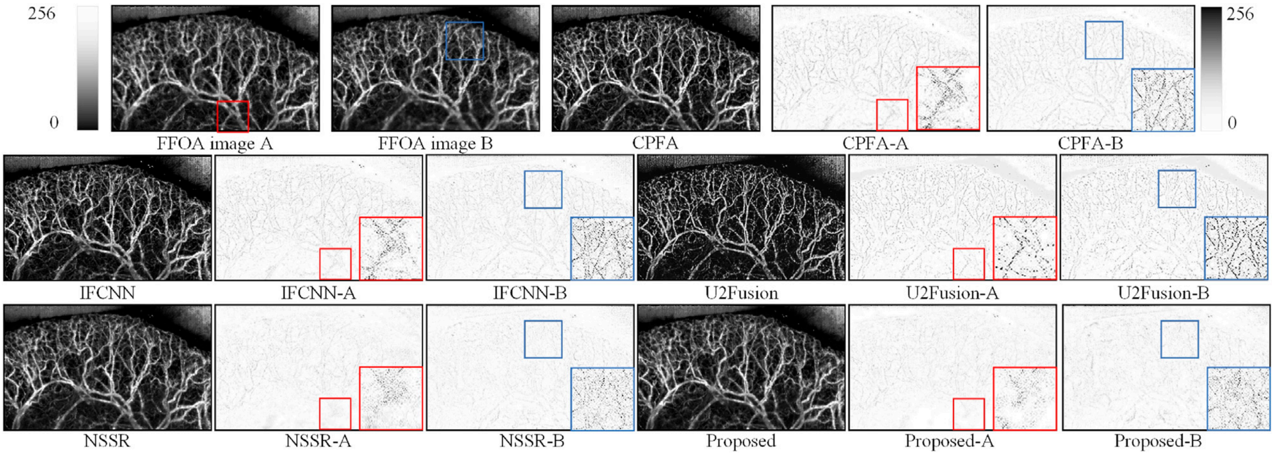


FIGURE 7
Subjective evaluation of the second group mouse ear experiments.

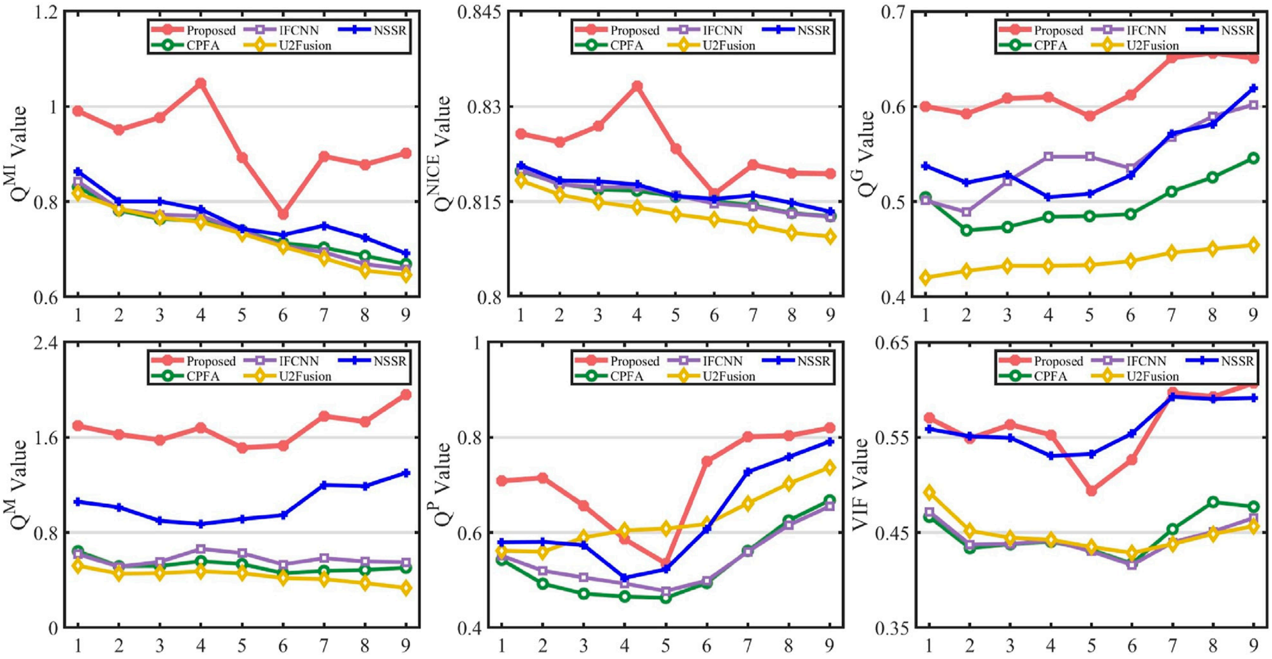


FIGURE 8
Objective evaluation of the first of group mouse ear.

images: the first image is the fused image, and the second and third images are the difference images. A comparison of the red-boxed regions shows that the residual information from the boxed regions of the proposed method and U2Fusion is smaller, which indicates that the GFD scheme was able to retain more information from the source image for different vessel thicknesses. Figure 7 shows the second group of mouse ear experiments. Here, the boxed region in the FFOA image A contains relatively more background tissue and fewer capillaries, and the boxed region in the FFOA image B contains rich capillary information; the other images in Figure 7 were obtained in the same manner as those in Figure 6. The blue zoomed area shows that there are cloud-like residuals in the fusion results of CPFA, IFCNN, and U2Fusion, suggesting that the GFD scheme can be effective for regions with fewer capillaries and more background tissue. In the difference images of the red focus region, CPFA, IFCNN, and U2Fusion show more evident vascular veins and lose some important contour edge details of the source images. NSSR also has a large number of residuals, demonstrating that NSSR poorly preserves the edge details of capillaries. The GFD scheme retains only a few residual information.

To evaluate the fusion results, Q_{MI}, Q_{NCIE}, Q_G, Q_M, Q_P, and VIF were used to evaluate nine fusions. Figures 8, 9 show the

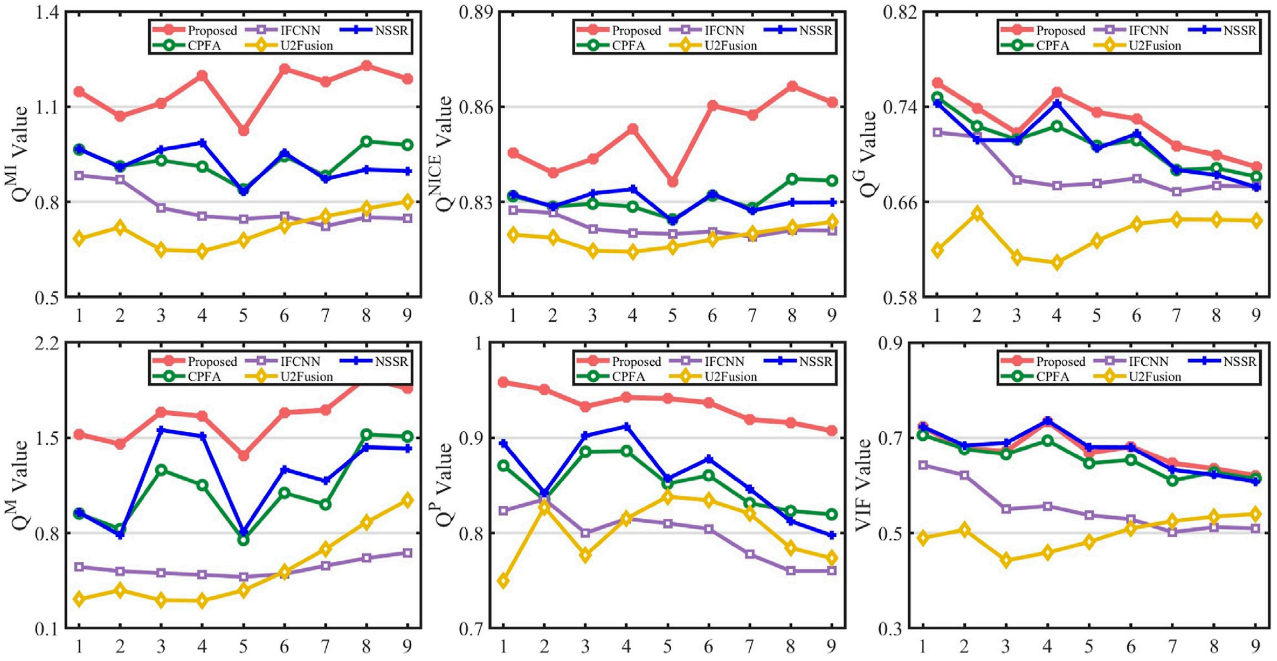


FIGURE 9
Objective evaluation of second of group mouse ear.

TABLE 1 Objective evaluation of mouse ears.

Experimental data	Method	Q_{MI}	Q_{NCIE}	Q_G	Q_M	Q_p	VIF
First set of images	CPFA	0.7385	0.8156	0.4945	0.5013	0.5195	0.4536
	IFCNN	0.7373	0.8159	0.5444	0.5774	0.5417	0.4435
	U2Fusion	0.7274	0.8133	0.4372	0.4333	0.6270	0.4486
	NSSR	0.7650	0.8160	0.5319	0.9749	0.6032	0.5588
	Proposed	**0.9231**	**0.8233**	**0.6189**	**1.6765**	**0.7082**	**0.5613**
Second set of images	CPFA	0.9282	0.8307	0.7092	1.1195	0.8515	0.6550
	IFCNN	0.7791	0.8219	0.6839	0.5273	0.7985	0.5513
	U2Fusion	0.7153	0.8186	0.6328	0.5331	0.8023	0.4987
	NSSR	0.9203	0.8301	0.7081	1.2126	0.8602	0.6727
	Proposed	**1.1517**	**0.8514**	**0.7254**	**1.6541**	**0.9338**	**0.6730**

The best results are in Bold.

metric values for the nine fusions of the ears of the first and second groups of mice, respectively. Table 1 lists the metric average values of the nine fusions; the optimal values are mentioned in bold font. Figures 8, 9 show that metrics Q_{MI}, Q_{NCIE}, Q_G, Q_M, and Q_P, than the other methods, and IFCNN and NSSR showed better performance, whereas CPFA and U2Fusion performed poorly. In terms of the VIF, the proposed method and NSSR showed excellent performance. Table 1 shows that the proposed method has the highest average value in terms of objective evaluation of the mouse ear in the first and second groups, and the NSSR also has good performance compared with other methods.

3.3 Fusion on the public dataset

To demonstrate the generalization of the proposed method, the Lytro dataset [64], which contains 20 pairs of multi-focus images, was used to validate the effectiveness of the method. The fusion results produced by the different methods on a set of Lytro dataset are shown in Figure 10. The average values of the objective evaluation of the Lytro dataset and Figure 10 are presented in Table 2.

In Figure 10, images A and B are produced by DOF in the same scene, which contains a motion field and a metal grid. The fusion result of each method consists of a fusion image and two difference images. The difference images were produced from the fusion result

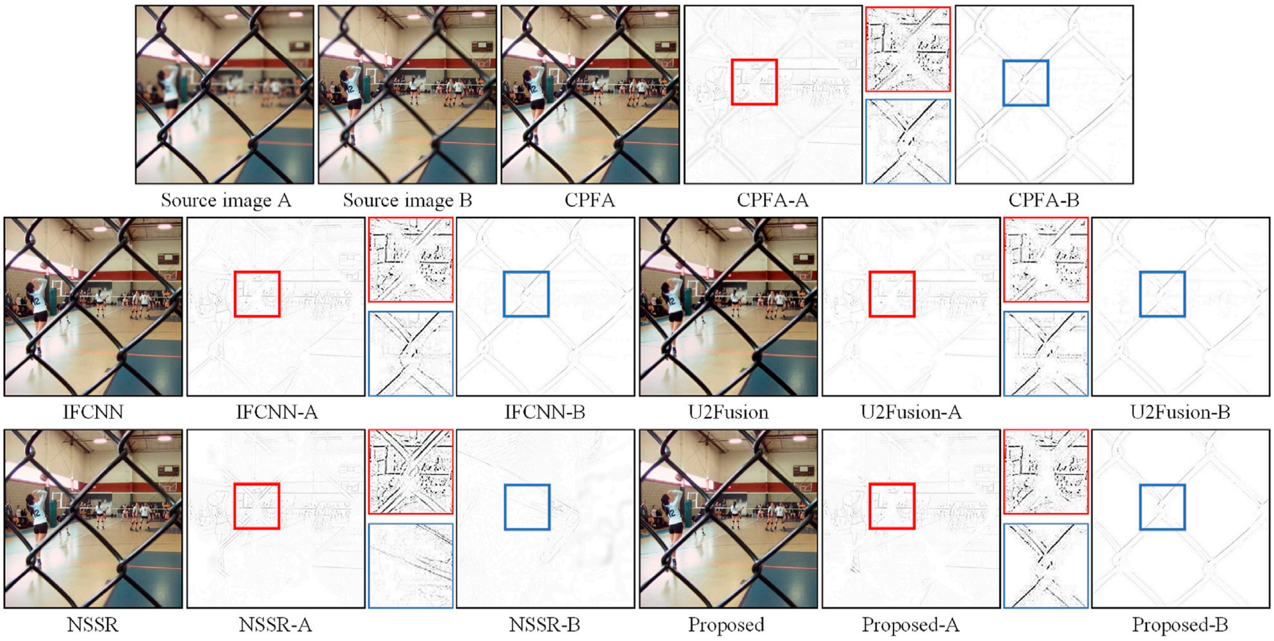


FIGURE 10
Subjective evaluation of the Lytro dataset.

TABLE 2 Objective evaluation of the Lytro dataset.

Experimental data	Method	Q_{MI}	Q_{NCIE}	Q_G	Q_M	Q_p	VIF
Objective evaluation of Figure 10	CPFA	0.9017	0.8306	0.6279	1.1729	0.8019	0.4652
	IFCNN	0.9545	0.8330	0.6377	0.8563	0.7984	0.4847
	U2Fusion	0.8144	0.8252	0.5614	0.5253	0.7474	0.4247
	NSSR	0.9289	0.8312	0.6216	1.6815	0.7788	**0.5234**
	Proposed	**1.0943**	**0.8437**	**0.6719**	**2.1190**	**0.8100**	0.5192
Average evaluation values of Lytro dataset	CPFA	0.9089	0.8286	0.6601	1.2671	0.8032	0.5086
	IFCNN	0.9377	0.8298	0.6628	0.9471	0.8178	0.5225
	U2Fusion	0.7989	0.8231	0.5601	0.4699	0.7272	0.4361
	NSSR	0.9493	0.8305	0.6869	1.7788	0.8183	0.5517
	Proposed	**1.1157**	**0.8406**	**0.7088**	**2.1405**	**0.8329**	**0.5634**

The best results are in Bold.

and original images A and B. Regions in the difference map containing focus and out-of-focus information were selected and enlarged in the middle of the two difference maps. From the overall fusion results, all methods can retain the brightness and color information in the source image satisfactorily; however, the fused images produced by NSSR and the proposed method achieve satisfactory results in terms of sharpness. The different methods showed a distinct gap in the difference images. In the difference images of CPFA, IFCCN, U2Fusion, and NSSR, residuals appeared in the focus region, indicating that these methods introduce information concerning the out-of-focus region in the fusion results. Particularly in the difference images NSSR-A and NSSR-B, there is almost no metal lattice shown in the out-of-focus images; this is attributable to the limited ability of the dictionary to characterize the image in the SR methods. In a comprehensive comparison, the proposed method achieved satisfactory subjective results in the subjective evaluation. In the objective evaluation results of the Lytro dataset, the proposed method achieved the best rankings in the metrics Q_{MI}, Q_{NCIE}, Q_G, Q_M and Q_P, although the value of the VIF was lower than that of NSSR, as shown in Figure 10. Considered together, the proposed method indicators were the best in the overall objective evaluation.

Based on the above discussion, this article confirmed the validity and stability of the proposed program. First, this is because, in contrast to CPFA, NSCT does not perform upsampling and downsampling. Thus, it reduces the redundancy between data in different layers and reduces the possibility of losing high-frequency detailed information in upsampling and downsampling, which may

TABLE 3 Running time of different methods.

Methods	CPFA	IFCNN	U2Fusion	NSSR	Proposed
Time/s	0.08	0.41	0.36	76.46	4.32

blur the fused images in the reconstruction process. Second, the NSCT can extract more accurate directional information to better represent image information. Finally, different fusion rules were adopted for different coefficients separately, which can stably retain the source image information. The proposed method could have potential applications in optical angiography experiments to extend the DOF.

3.4 Discussion on time efficiency

In this section, the time efficiency of the proposed method will be compared with other methods on grayscale images (size 710 × 620). As summarized in Table 3, the NSSR method takes the longest time because it uses a dictionary for the SR of the image. In contrast, the CPFA has the shortest time because of the fast contrast pyramid construction process and the simple fusion rules used. The computational efficiencies of the deep learning methods IFCNN and U2Fusion were relatively high because they use pre-trained models. In terms of the time required, proposed method ranked fourth; this is attributable to the large amount of time spent on the NSCT decomposition and the relative complexity of the computation of the fusion rule. The speed of proposed method may not be the highest, but its high performance makes it effective. Additionally, optimizing the underlying code and utilizing tools such as GPUs and C++ holds the potential to significantly reduce the execution time of proposed method, which will enable the method to meet the requirements of a wider range of applications.

4 Conclusion

Blood microcirculation information is essential for biological research. This article developed a GFD method to solve the defocusing problems by extending the DOF. FFOA images with different DOFs were obtained using the AIFM effect; subsequently, the DOF was extended using the proposed fusion method. The proposed fusion methodology consists of three steps. First, the NSCT decomposes the FFOA images into LFC and HFDCs. GFD rules are employed to fuse the LFC and HFDCs, and the final fused images are obtained by performing INSCT. Subjective visual comparison and objective assessment in the experiments can certify the validity and stability of the proposed scheme. Experimental results show that the proposed method can solve the FFOA scattering problem biological samples due to surface and thickness inhomogeneity, and has the potential applications in optical angiography experiments; notably, it provides effective technical support for target identification and tracking.

Although the proposed GFD method can obtain high-resolution blood flow images by extending the DOF, there are some limitations. First, the EZL has a limited focusing speed, resulting in the inability to image in real time. Second, the decomposition level of NSP and decomposition direction of NSDFB in the NSCT must be set using artificial empirical values, which increases the uncertainty of the fusion effect; moreover, the computational efficiency of the GFD needs to be refined. Finally, the completed FFOA image must be registered to reduce artifacts from the sample jitter in the fused image. In future work, the designed algorithm will be improved to enhance the robustness of fusing noise-disturbing and misregistered images.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The animal study was approved by School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528225, China. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

GW: Visualization, Writing–original draft. JL: Conceptualization, Methodology, Software, Writing–review and editing. HT: Data curation, Supervision, Writing–review and editing. XL: Funding acquisition, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research is funded by the National Natural Science Foundation of China under Grant (No. 62201149), the Joint Fund for Basic and Applied Basic Research of Guangdong Province under Grant (No. 2023A1515140077), the Natural Science Foundation of Guangdong Province under Grant (No. 2024A1515011880), the Guangdong Higher Education Innovation and Strengthening of Universities Project under Grant (No. 2023KTSCX127), and the Foshan Key Areas of Scientific and Technological Research Project under Grant (No. 2120001008558).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Taqueti VR, Di Carli MF Coronary microvascular disease pathogenic mechanisms and therapeutic options JACC state-of-the-art review. *J Am Coll Cardiol* (2018) 72(21): 2625–41. doi:10.1016/j.jacc.2018.09.042

2. Simo R, Stitt AW, Gardner TW. Neurodegeneration in diabetic retinopathy: does it really matter? *Diabetologia* (2018) 61(9):1902–12. doi:10.1007/s00125-018-4692-1

3. Montone RA, Niccoli G, Fracassi F, Russo M, Gurgoglione F, Cammà G, et al. Patients with acute myocardial infarction and non-obstructive coronary arteries: safety and prognostic relevance of invasive coronary provocative tests. *Eur Heart J* (2018) 39(2):91–8. doi:10.1093/eurheartj/ehx667

4. Feihl F, Liaudet L, Waeber B, Levy BI. Hypertension - a disease of the microcirculation? *Hypertension* (2006) 48(6):1012–7. doi:10.1161/01.hyp.0000249510.20326.72

5. de Boer JF, Hitzenberger CK, Yasuno Y. Polarization sensitive optical coherence tomography - a review Invited. *Biomed Opt Express* (2017) 8(3):1838–73. doi:10.1364/boe.8.001838

6. Briers D, Duncan DD, Hirst E, Kirkpatrick SJ, Larsson M, Steenbergen W, et al. Laser speckle contrast imaging: theoretical and practical limitations. *J Biomed Opt* (2013) 18(6):066018. doi:10.1117/1.jbo.18.6.066018

7. Zhang FL, Wang MY, Han DA, Tan H, Yang G, Zeng Y. *In vivo* full-field functional optical hemocytometer. *J Biophotonics* (2018) 11(2). doi:10.1002/jbio.201700039

8. Wang MY, Mao WJ, Guan CZ, Feng G, Tan H, Han D, et al. Full-field functional optical angiography. *Opt Lett* (2017) 42(3):635–8. doi:10.1364/ol.42.000635

9. Liu Y, Wang L, Cheng J, Chen X. Multi-focus image fusion: a Survey of the state of the art. *Inf Fusion* (2020) 64:71–91. doi:10.1016/j.inffus.2020.06.013

10. Zhu ZQ, Zheng MG, Qi GQ, Wang D, Xiang Y. A Phase congruency and local laplacian energy based multi-modality medical image fusion method in NSCT domain. *Ieee Access* (2019) 7:20811–24. doi:10.1109/access.2019.2898111

11. Li XS, Zhou FQ, Tan HS, et al. Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal. *Signal Process.* (2021) 184. doi:10.1016/j.sigpro.2021.108062

12. Meher B, Agrawal S, Panda R, Abraham A. A survey on region based image fusion methods. *Inf Fusion* (2019) 48:119–32. doi:10.1016/j.inffus.2018.07.010

13. Li X, Li X, Ye T, Cheng X, Liu W, Tan H. Bridging the gap between multi-focus and multi-modal: a focused integration framework for multi-modal image fusion. *arXiv preprint arXiv:231101886,* (2024):2023. doi:10.1109/wacv57701.2024.00165

14. Li J, Han D, Wang X, Yi P, Yan L, Li X. Multi-Sensor medical-image fusion technique based on embedding bilateral filter in least squares and salient detection. *Sensors* (2023) 23(7):3490. doi:10.3390/s23073490

15. Wang JW, Qu HJ, Wei YA, et al. Multi-focus image fusion based on quad-tree decomposition and e dge-weighte d focus measure. *Signal Process.* (2022) 198. doi:10.1016/j.sigpro.2022.108590

16. Li ST, Yang B, Hu JW. Performance comparison of different multi-resolution transforms for image fusion. *Inf Fusion* (2011) 12(2):74–84. doi:10.1016/j.inffus.2010.03.002

17. Li XL, Wang XP, Cheng XQ, Tan H. Multi-focus image fusion based on hessian matrix decomposition and salient difference focus detection. *Entropy* (2022) 24(11): 1527. doi:10.3390/e24111527

18. Li XS, Zhou FQ, Tan HS, Zhang W, Zhao C. Multimodal medical image fusion based on joint bilateral filter and local gradient energy. *Inf Sci* (2021) 569:302–25. doi:10.1016/j.ins.2021.04.052

19. Jie YC, Li XS, Wang MY, et al. Medical image fusion based on extended difference-of-Gaussians and edge-preserving. *Expert Syst Appl* (2023) 227. doi:10.1016/j.eswa.2023.120301

20. Zhu ZQ, Yin HP, Chai Y, Li Y, Qi G. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inf Sci* (2018) 432:516–29. doi:10.1016/j.ins.2017.09.010

21. Zhang YF, Yang MY, Li N, et al. Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion. *Signal Process.* (2020) 167. doi:10.1016/j.sigpro.2019.107327

22. Li HF, Wang YT, Yang Z, Wang R, Li X, Tao D. Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *Ieee Trans Instrumentation Meas* (2020) 69(4):1082–102. doi:10.1109/tim.2019.2912239

23. Li HF, He XG, Tao DP, Tang Y, Wang R. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognition* (2018) 79:130–46. doi:10.1016/j.patcog.2018.02.005

24. Li XS, Zhou FQ, Tan HS. Joint image fusion and denoising via three-layer decomposition and sparse representation. *Knowledge-Based Syst* (2021) 224. doi:10.1016/j.knosys.2021.107087

25. Li XS, Wan WJ, Zhou FQ, Cheng X, Jie Y, Tan H. Medical image fusion based on sparse representation and neighbor energy activity. *Biomed Signal Process Control* (2023) 80:104353. doi:10.1016/j.bspc.2022.104353

26. Liu Y, Chen X, Peng H, Wang Z. Multi-focus image fusion with a deep convolutional neural network. *Inf Fusion* (2017) 36:191–207. doi:10.1016/j.inffus.2016.12.001

27. Ma JY, Tang LF, Fan F, Huang J, Mei X, Ma Y. SwinFusion: cross-domain long-range learning for general image fusion via swin transformer. *Ieee-Caa J Automatica Sinica* (2022) 9(7):1200–17. doi:10.1109/jas.2022.105686

28. Zhang XC. Deep learning-based multi-focus image fusion: a survey and a comparative study. *Ieee Trans Pattern Anal Machine Intelligence* (2022) 44(9): 4819–38. doi:10.1109/tpami.2021.3078906

29. Liu Y, Qi Z, Cheng J, Chen X. Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: a statistic-based approach. *IEEE Trans pattern Anal machine intelligence* (2024) 1–14. doi:10.1109/tpami.2024.3367905

30. Li HF, Liu JY, Zhang YF, Liu Y. A deep learning framework for infrared and visible image fusion without strict registration. *Int J Comput Vis* (2023). doi:10.1007/s11263-023-01948-x

31. Xiao B, Ou G, Tang H, Bi X, Li W. Multi-focus image fusion by hessian matrix based decomposition. *Ieee Trans Multimedia* (2020) 22(2):285–97. doi:10.1109/tmm.2019.2928516

32. Li X, Li X, Tan H, Li J. SAMF: small-area-aware multi-focus image fusion for object detection. *arXiv preprint arXiv:240108357* (2024) doi:10.1109/icassp48485.2024.10447642

33. Li XL, Li XS, Cheng XQ, Wang M, Tan H. MCDFD: multifocus image fusion based on multiscale cross-difference and focus detection. *IEEE Sens J* (2023) 23(24):30913–26. doi:10.1109/jsen.2023.3330871

34. Zhang Q, Liu Y, Blum RS, Han J, Tao D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Inf Fusion* (2018) 40: 57–75. doi:10.1016/j.inffus.2017.05.006

35. Tang D, Xiong QY, Yin HP, et al. A novel sparse representation based fusion approach for multi-focus images. *Expert Syst Appl* (2022) 197. doi:10.1016/j.eswa.2022.116737

36. Guo XP, Nie RC, Cao JD, Zhou D, Mei L, He K. FuseGAN: learning to fuse multi-focus image via conditional generative adversarial network. *Ieee Trans Multimedia* (2019) 21(8):1982–96. doi:10.1109/tmm.2019.2895292

37. Luo X, Gao Y, Wang A, et al. IFSepR: a general framework for image fusion based on separate representation learning. *IEEE Trans Multimedia* (2021) 1. doi:10.1109/TMM.2021.3129354

38. Zhu Z, Sun M, Qi G, Li Y, Gao X, Liu Y. Sparse Dynamic Volume TransUNet with multi-level edge fusion for brain tumor segmentation. *Comput Biol Med* (2024) 172: 108284. doi:10.1016/j.compbiomed.2024.108284

39. Li J, Huo H, Li C, Wang R, Feng Q. AttentionFGAN: infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans Multimedia* (2020) 23:1383–96. doi:10.1109/tmm.2020.2997127

40. Tang W, He F, Liu Y, Duan Y. MATR: multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans Image Process* (2022) 31:5134–49. doi:10.1109/tip.2022.3193288

41. Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, et al. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. *Computer Vis Pattern Recognition*, 2023: 5906–16. doi:10.1109/CVPR52729.2023.00572

42. Zhao HJ, Shang ZW, Tang YY, Fang B. Multi-focus image fusion based on the neighbor distance. *Pattern Recognition* (2013) 46(3):1002–11. doi:10.1016/j.patcog.2012.09.012

43. Petrovic VS, Xydeas CS. Gradient-based multiresolution image fusion. *Ieee Trans Image Process* (2004) 13(2):228–37. doi:10.1109/tip.2004.823821

44. Wan T, Canagarajah N, Achim A. Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients. *Ieee Trans Multimedia* (2009) 11(4): 624–33. doi:10.1109/TMM.2009.2017640

45. Lewis JJ, O'Callaghan RJ, Nikolov SG, et al. Pixel- and region-based image fusion with complex wavelets. *Inf Fusion* (2007) 8(2):119–30. doi:10.1016/j.inffus.2005.09.006

46. Zhang Q, Guo BL. Multifocus image fusion using the nonsubsampled contourlet transform. *Signal Process.* (2009) 89(7):1334–46. doi:10.1016/j.sigpro.2009.01.012

47. da Cunha AL, Zhou JP, Do MN. The nonsubsampled contourlet transform: theory, design, and applications. *Ieee Trans Image Process* (2006) 15(10):3089–101. doi:10.1109/tip.2006.877507

48. Wang MY, Wu NS, Huang HH, Luo J, Zeng Y, et al. Large-depth-of-field full-field optical angiography. *J Biophotonics* (2019) 12(5):e201800329. doi:10.1002/jbio.201800329

49. Huang W, Jing Z. Evaluation of focus measures in multi-focus image fusion. *Pattern Recognition Lett* (2007) 28(4):493–500. doi:10.1016/j.patrec.2006.09.005

50. Zhu X, Milanfar P. Automatic parameter selection for denoising algorithms using a No-reference measure of image content. *Ieee Trans Image Process* (2010) 19(12): 3116–32. doi:10.1109/TIP.2010.2052820

51. Du J, Li WS, Tan HL. Three-layer medical image fusion with tensor-based features. *Inf Sci* (2020) 525:93–108. doi:10.1016/j.ins.2020.03.051

52. Jin L, Liu H, Xu X, Song E. Improved direction estimation for Di Zenzo's multichannel image gradient operator. *Pattern Recognition* (2012) 45(12):4300–11. doi:10.1016/j.patcog.2012.06.003

53. Zhou Z, Li S, Wang B. Multi-scale weighted gradient-based fusion for multi-focus images. *Inf Fusion* (2014) 20:60–72. doi:10.1016/j.inffus.2013.11.005

54. Qu GH, Zhang DL, Yan PF. Information measure for performance of image fusion. *Electron Lett* (2002) 38(7):313–5. doi:10.1049/el:20020212

55. Wang Q, Shen Y, Zhang JQ. A nonlinear correlation measure for multivariable data set. *Physica D-Nonlinear Phenomena* (2005) 200(3-4):287–95. doi:10.1016/j.physd.2004.11.001

56. Xydeas CS, Petrovic V. Objective image fusion performance measure. *Electron Lett* (2000) 36(4):308–9. doi:10.1049/el:20000267

57. Wang PW, Liu B (2008). "A novel image fusion metric based on multi-scale analysis," in 9th international conference on signal processing, 26-29 October 2008, beijing.

58. Zhao JY, Laganiere R, Liu Z. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int J Innovative Comput Inf Control* (2007) 3(6A):1433–47.

59. Sheikh HR, Bovik AC. Image information and visual quality. *Ieee Trans Image Process* (2006) 15(2):430–44. doi:10.1109/tip.2005.859378

60. Liu Z, Blasch E, Xue ZY, Zhao J, Laganiere R, Wu W. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. *Ieee Trans Pattern Anal Machine Intelligence* (2012) 34(1):94–109. doi:10.1109/tpami.2011.109

61. Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L. IFCNN: a general image fusion framework based on convolutional neural network. *Inf Fusion* (2020) 54:99–118. doi:10.1016/j.inffus.2019.07.011

62. Xu H, Ma JY, Jiang JJ, Guo X, Ling H. U2Fusion: a unified unsupervised image fusion network. *Ieee Trans Pattern Anal Machine Intelligence* (2022) 44(1):502–18. doi:10.1109/tpami.2020.3012548

63. Jie YC, Li XS, Wang MY, Tan H. Multi-focus image fusion for full-field optical angiography. *Entropy* (2023) 25(6):951. doi:10.3390/e25060951

64. Nejati M, Samavi S, Shirani S. Multi-focus image fusion using dictionary-based sparse representation. *Inf Fusion* (2015) 25:72–84. doi:10.1016/j.inffus.2014.10.004

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 15 May 2024
DOI 10.3389/fphy.2024.1392767

Check for updates

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Zhongyuan Zhao,
Nanjing University of Information Science and Technology, China
Pan Yang,
Chongqing Medical University, China

*CORRESPONDENCE
Ke Wang,
✉ w13996048216@163.com
Sheng Chen,
✉ up_chen@126.com

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 28 February 2024
ACCEPTED 26 April 2024
PUBLISHED 15 May 2024

CITATION
Huang J, Chen F, Wang K and Chen S (2024), Research on electrical capacitance tomography (ECT) detection of cerebral hemorrhage based on symmetrical cancellation method.
Front. Phys. 12:1392767.
doi: 10.3389/fphy.2024.1392767

Research on electrical capacitance tomography (ECT) detection of cerebral hemorrhage based on symmetrical cancellation method

Jing Huang[1†], Feng Chen[2†], Ke Wang[3]* and Sheng Chen[2]*

[1]Department of Neurology, First Hospital, Shanxi Medical University, Taiyuan, China, [2]Chongqing University Central Hospital, Chongqing Emergency Medical Center, Chongqing, China, [3]Department of Neurosurgery, Chongqing University Central Hospital, Chongqing Emergency Medical Center, Chongqing, China

Currently, there is an urgent need for a fast and portable intracerebral hemorrhage (ICH) detection technology for pre-hospital emergency scenarios. Owing to the disproportionately elevated permittivity of blood compared to other brain tissues, Electrical Capacitance Tomography (ECT) offers a viable modality for mapping the spatial distribution of permittivity within the brain, thus facilitating the imaging-based identification of ICH. Currently, ECT is confined to time-differential imaging due to limited sensitivity, and this methodology requires non-hemorrhagic measurements for comparison, data that are frequently inaccessible in clinical contexts. To overcome this limitation, in accordance with the natural bilateral symmetry of the cerebral hemispheres, a symmetrical cancellation scheme is introduced. In this method, electrodes are uniformly arrayed around the cranial periphery and strategically positioned in a symmetrical manner relative to the sagittal suture. Subsequently, the measured capacitances for each electrode pair are subtracted from those of their symmetrical counterparts aligned with the sagittal suture. As a result, this process isolates the capacitance attributable solely to hemorrhagic events within a given hemisphere, permitting the absolute imaging of ICH. To assess the feasibility of this method, simulation and empirical imaging were conducted respectively on a numerical hemorrhage model and three physical models (a water-wrapped hemorrhage model, an isolated porcine fat-wrapped hemorrhage model, and an isolated porcine brain tissue-wrapped hemorrhage model). Traditional absolute imaging, time-differential imaging and symmetrical cancellation imaging were performed on all models. The results substantiate that the proposed imaging modality is capable of obtaining absolute imaging of ICH. But a mirrored artifact, symmetrical to the site of the actual hemorrhage image appeared in each of the imaging results. This mirror artifact was characterized by identical dimensions and an inverted pixel-value schema, an intrinsic consequence of the symmetrical cancellation imaging algorithm. The real image of hemorrhage can be ascertained through pre-judgment with the symptoms of the patient. Additionally, the quality of this imaging is seriously dependent on the precise alignment between the electrodes and the sagittal suture of the brain; even a minor deviation in symmetry could introduce excessive noises. Thus, the complicated operational procedures remain as challenges for practical application.

KEYWORDS

intracerebral hemorrhage, ECT, symmetrical cancellation imaging, prehospital emergency, measurement

Introduction

Spontaneous intracerebral hemorrhage (ICH) constitutes hemorrhage induced by the disruption of blood vessels within the brain parenchyma. It represents the most grave form of acute stroke due to its immediacy, perilous nature, high morbidity, and mortality. Annually, ICH is accountable for approximately 2.8 million fatalities, yielding an incidence rate of 4.1% [1]. According to the 2018 China Stroke Prevention and Treatment Report, the incidence rate of hemorrhagic stroke was 126.34 per 100,000 person-years in China [2]. Optimal postoperative outcomes and survival rates subsequent to ICH could be substantially improved through prompt diagnosis and intervention [3, 4]. Presently, CT and MRI scans constitute the primary modalities for ICH detection. However, considerable temporal lags ensue during patient transit to healthcare facilities, CT examination, and consequent diagnostic revelation, thereby losing the most ideal time for effective treatment. Moreover, these voluminous pieces of equipment are infeasible for pre-hospital emergency care and bedside monitoring. Hence, a portable, cost-effective, non-invasive, and expedient detection technology for ICH is imperatively necessitated.

Innovative methodologies, aimed at diagnosing cerebral pathologies, capitalize on the electrical properties of biological tissues, notably exemplified by Electrical Impedance Tomography (EIT) and Magnetic Induction Tomography (MIT) [5, 6]. Due to the relatively high electrical impedance of the skull, there is a considerable attenuation of the exciting current in EIT. Secondly, EIT requires the connection of electrodes with the scalp, which results in a very large contact impedance. These issues result in low sensitivity of EIT in imaging brain tissues. As for MIT, the induced magnetic field generated in biological tissues exposed to an excitation field is negligible because of the biological tissue’s poor conductivity (0.1–2 S/m). Furthermore, the conductivity of blood is not noticeably different from those of other brain tissues. Because of these two factors, MIT has relatively poor sensitivity for visualizing a brain hemorrhage.

Investigations into the permittivity of cerebral tissues have elucidated that the permittivity of blood markedly supersedes that of other tissues. At a frequency of 1 MHz, the permittivity values for blood, grey matter, and cerebrospinal fluid stand at 3,000, 990, and 108, respectively [7]. Albeit the permittivity across all cerebral tissues diminishes in tandem with frequency, the permittivity indices of blood remain uniformly elevated. Hence, theoretical considerations suggest that imaging based on permittivity distributions is more efficacious than conductivity-based imaging for ICH detection. Electrical Capacitance Tomography (ECT) serves as a technological platform for visualizing permittivity distribution within the object under examine, predicated upon capacitance measurements obtained from a multi-electrode sensor encircling said object—a technique commonly employed in multi-phase flow analyses in the oil sector and fluidized bed measurements in the industry [8, 9]. Previous experimental endeavors have utilized parallel plate capacitors to measure cerebral capacitance change concurrent with hemorrhagic events; results from animal studies showed an increment in cerebral capacitance change with increased volumes of blood infusion [10]. Subsequently, we engineered a 16-channel ECT system, successfully employing it to visualize hemorrhagic phenomena within porcine cerebral tissue *ex vivo* [11]. These preliminary studies proved the feasibility of detecting the onset of cerebral hemorrhage via capacitance variations in brain tissue. Although the utility of ECT in our last study for *in vitro* imaging of cerebral hemorrhage, the employed methodology was that of time-differential imaging—subtracting pre-hemorrhagic measurement data from post-hemorrhagic data—which is commonly adopted in most current electrical imaging modalities [12]. Given that baseline, non-hemorrhagic data is unattainable in clinical settings, this approach is restricted to the dynamic monitoring of bleeding, thus lacking the ability for initial ICH diagnosis. To fulfill the unmet need for immediate ICH detection, it is imperative to ascertain the absolute spatial distribution of cerebral hemorrhage, rather than its temporal change, akin to the capabilities of CT and MRI. Owing to the subtle electrical property differentials between cerebral hemorrhage and other biological tissues, coupled with the minuscule volume of cerebral hemorrhage relative to normative cerebral tissues, the weak signal emanating from the hemorrhagic region is subsumed within the electrical noise generated by normal cerebral tissue. Consequently, conventional electrical imaging techniques are incapable of delineating the absolute electrical parameter distributions within the entire cerebral domain, much less those specifically related to cerebral hemorrhage.

The examination of the structural composition of the human brain demonstrates that the left and right hemispheres are substantially symmetrical with respect to the sagittal suture, and the histological distributions within these hemispheres are analogous [13, 14]. Numerous studies indicate that the impedance in the left and right hemispheres of a healthy brain is comparatively homogenous. Empirical evidence from numerous cases reveals that most cerebral hemorrhages (excluding subarachnoid hemorrhages) manifest in a single hemisphere, and localized hemorrhagic events do not perturb the tissue distribution of the contralateral hemisphere in the absence of a midline shift [15, 16]. However, the occurrence of localized bleeding within a hemisphere disrupts the impedance equilibrium between the two hemispheres [17]. In this study, a modified ECT imaging—termed symmetrical cancellation ECT—is proffered, predicated upon the inherent structural characteristics of the human brain. In this

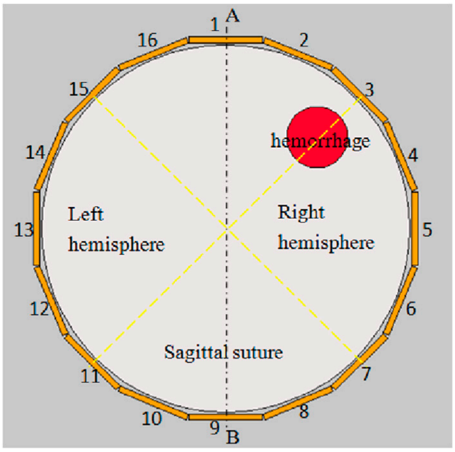


FIGURE 1
The schematic diagram of ECT Sensor with symmetrical cancellation.

paradigm, electrodes strategically arrayed around the cranium are positioned in a symmetrical manner relative to the sagittal suture. Subsequently, the measured capacitance from each electrode pair is subtracted from the reference capacitance of the symmetrical electrode pair adjacent to the sagittal suture, serving as the imaging data. This imaging technique is devised to counteract all capacitance signals emanating from normative brain tissue at symmetrical positions within the left and right hemispheres, isolating only the capacitance signals attributable to hemorrhagic events. In theory, this enables the acquisition of absolute imaging of cerebral hemorrhage [18]. Barry McDermott et al. [19] applied this approach in EIT, arranging the EIT electrodes symmetrically on both hemispheres of the skull and using the voltage differences between symmetrically placed electrodes as imaging data. In their simulations and physical experiments, they achieved absolute imaging of cerebral hemorrhage. The imaging results included not only the hemorrhage image but also a mirrored image with pixel values that were inversely related. Subsequently, they proposed a dual-frequency symmetrical cancellation EIT imaging method to minimize the impact of electrode symmetry errors on the imaging outcomes [20]. These two articles verify the feasibility of this method.

This paper substantiates the feasibility of the proposed method via simulation and empirical imaging exercises. In the simulation experiment, a cerebral hemorrhage model comprising six distinct tissue types was constructed, and cerebral hemorrhage was depicted using both time-differential imaging and symmetrical cancellation imaging. In the experimental phase, three prototypical cerebral hemorrhage models were established: a model of water-encapsulated blood, a model of isolated porcine adipose tissue-encapsulated blood, and a model with isolated porcine cerebral tissue-encapsulated blood. Subsequent imaging utilizing the aforementioned 16-electrode ECT system was conducted on these models, employing absolute imaging, time-differential imaging, and symmetrical cancellation imaging. The resultant imaging outcomes were then comparatively analyzed.

Methods and materials

Symmetrical cancellation ECT method

The typical ECT imaging system is composed of three components: 1) sensor, 2) data acquisition system, and 3) computer for reconstruction. Figure 1 illustrates an ECT sensor equipped with 16 electrodes, denoted by integers 1 to 16. These 16 homogeneous electrodes are uniformly arrayed around the cranium (represented by a sizable white circle at the center). The spherical cranium is partitioned into the left and right hemispheres, and exhibits symmetry about the central sagittal suture (delineated by dashed line AB). A crimson circle is situated in the upper-right quadrant of the right hemisphere to signify hemorrhage. To implement a symmetrical cancellation measurement, it is important that the electrodes dispersed in the left and right hemispheres maintain symmetry relative to the sagittal suture. To satisfy this criterion, one must ensure that the sagittal suture of the skull and the midpoint of a pair of opposing electrodes are collinear. In Figure 1, the sagittal suture is aligned with the midpoint of electrodes 1 and 9; that is, dashed line AB intersects the centers of electrodes 1 and 9 as well as the central imaging region. Consequently, electrodes 2 and 16 are symmetrical, as are electrodes 3 and 15, electrodes 4 and 14, electrodes 5 and 13, electrodes 6 and 12, electrodes 7 and 11, and electrodes 8 and 10. AB additionally serves as the axis of symmetry for the ECT sensor. In a complete measurement protocol within conventional ECT systems, an electrode is successively chosen as the excitation electrode, while the remaining serve as detection electrodes, to obtain the capacitance data between all different electrode pairs. Taking an sixteen-electrode sensor in Figure 1 as an example, capacitance measurements are obtained in the following sequential steps. Initially, a voltage signal is administered to electrode 1, followed by the measurement of electric charges on electrodes 2–16, thereby quantifying the capacitances between electrode 1 and the other 15 electrodes. Subsequently, electrodes 2–15 are activated in a systematic sequence, thus enabling the acquisition of capacitance data for all unique electrode pairs, culminating in a total of 120 independent electrode combinations. With this measurement strategy, the number of independent capacitance measurements is

$$M = N\,(N - 1)/2 \tag{1}$$

where N is the number of electrodes. For this particular sensor, $N = 16$, and 120 independent capacitances can be measured from different electrode pairs.

ECT imaging data are typically rendered by subtracting the reference frame data from the substance-field measurements. Various imaging modalities employ different sets of reference frame data. In traditional absolute imaging, the reference data are the measurement data (i.e., the empty field measurement data) when the imaging area is entirely comprised of air. For time-differential imaging, the reference data are derived from pre-hemorrhagic measurements; however, it is impossible to acquire such measurements in a real-world setting. The reference data for each electrode pair in symmetrical cancellation ECT proposed herein utilizes the data of the electrode pair exhibiting axial symmetry,

TABLE 1 All the measuring electrode pairs and their corresponding reference electrode pairs in symmetrical cancellation ECT when electrodes 1, 5, 9, 13 are used as excitation electrodes.

Excitation electrode 1		Excitation electrode 5		Excitation electrode 9		Excitation electrode 13	
Measuring electrode pairs	Reference electrode pairs	Measuring electrode pairs	Reference electrode pairs	Measuring electrode pairs	Reference electrode pairs	Measuring electrode pairs	Reference electrode pairs
$C_{1\text{-}2}$	$C_{1\text{-}16}$	$C_{5\text{-}6}$	$C_{13\text{-}12}$	$C_{9\text{-}10}$	$C_{9\text{-}8}$	$C_{13\text{-}14}$	$C_{5\text{-}4}$
$C_{1\text{-}3}$	$C_{1\text{-}15}$	$C_{5\text{-}7}$	$C_{13\text{-}11}$	$C_{9\text{-}11}$	$C_{9\text{-}7}$	$C_{13\text{-}15}$	$C_{5\text{-}3}$
$C_{1\text{-}4}$	$C_{1\text{-}14}$	$C_{5\text{-}8}$	$C_{13\text{-}10}$	$C_{9\text{-}12}$	$C_{9\text{-}6}$	$C_{13\text{-}16}$	$C_{5\text{-}2}$
$C_{1\text{-}5}$	$C_{1\text{-}13}$	$C_{5\text{-}9}$	$C_{13\text{-}9}$	$C_{9\text{-}13}$	$C_{9\text{-}5}$	$C_{13\text{-}1}$	$C_{5\text{-}1}$
$C_{1\text{-}6}$	$C_{1\text{-}12}$	$C_{5\text{-}10}$	$C_{13\text{-}8}$	$C_{9\text{-}14}$	$C_{9\text{-}4}$	$C_{13\text{-}2}$	$C_{5\text{-}16}$
$C_{1\text{-}7}$	$C_{1\text{-}11}$	$C_{5\text{-}11}$	$C_{13\text{-}7}$	$C_{9\text{-}15}$	$C_{9\text{-}3}$	$C_{13\text{-}3}$	$C_{5\text{-}15}$
$C_{1\text{-}8}$	$C_{1\text{-}10}$	$C_{5\text{-}12}$	$C_{13\text{-}6}$	$C_{9\text{-}16}$	$C_{9\text{-}2}$	$C_{13\text{-}4}$	$C_{5\text{-}14}$
$C_{1\text{-}9}$	$C_{1\text{-}9}$	$C_{5\text{-}13}$	$C_{13\text{-}5}$	$C_{9\text{-}1}$	$C_{9\text{-}1}$	$C_{13\text{-}5}$	$C_{5\text{-}13}$
$C_{1\text{-}10}$	$C_{1\text{-}8}$	$C_{5\text{-}14}$	$C_{13\text{-}4}$	$C_{9\text{-}2}$	$C_{9\text{-}16}$	$C_{13\text{-}6}$	$C_{5\text{-}12}$
$C_{1\text{-}11}$	$C_{1\text{-}7}$	$C_{5\text{-}15}$	$C_{13\text{-}3}$	$C_{9\text{-}3}$	$C_{9\text{-}15}$	$C_{13\text{-}7}$	$C_{5\text{-}11}$
$C_{1\text{-}12}$	$C_{1\text{-}6}$	$C_{5\text{-}16}$	$C_{13\text{-}2}$	$C_{9\text{-}4}$	$C_{9\text{-}14}$	$C_{13\text{-}8}$	$C_{5\text{-}10}$
$C_{1\text{-}13}$	$C_{1\text{-}5}$	$C_{5\text{-}1}$	$C_{13\text{-}1}$	$C_{9\text{-}5}$	$C_{9\text{-}13}$	$C_{13\text{-}9}$	$C_{5\text{-}9}$
$C_{1\text{-}14}$	$C_{1\text{-}4}$	$C_{5\text{-}2}$	$C_{13\text{-}16}$	$C_{9\text{-}6}$	$C_{9\text{-}12}$	$C_{13\text{-}10}$	$C_{5\text{-}8}$
$C_{1\text{-}15}$	$C_{1\text{-}3}$	$C_{5\text{-}3}$	$C_{13\text{-}15}$	$C_{9\text{-}7}$	$C_{9\text{-}11}$	$C_{13\text{-}11}$	$C_{5\text{-}7}$
$C_{1\text{-}16}$	$C_{1\text{-}2}$	$C_{5\text{-}4}$	$C_{13\text{-}14}$	$C_{9\text{-}8}$	$C_{9\text{-}10}$	$C_{13\text{-}12}$	$C_{5\text{-}6}$

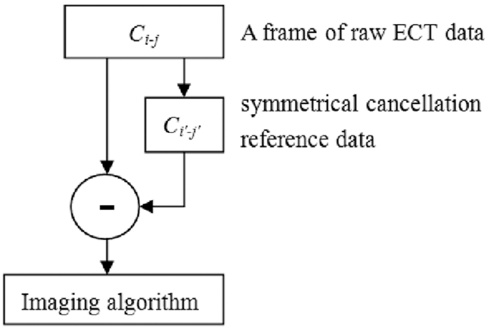


FIGURE 2
The flow chart of symmetrical cancellation ECT.

therefore the measurement data and the reference data are all derived from the substance-field measurement data. $C_{i\text{-}j}$ is utilized to denote the capacitance of the electrode pair comprising the excitation electrode i and the measurement electrode j. As depicted in Figure 1, for instance, with electrode 1 serving as the excitation electrode, the reference electrode pair for $C_{1\text{-}2}$ is $C_{1\text{-}16}$, the reference electrode pair for $C_{1\text{-}3}$ is $C_{1\text{-}15}$, . . ., the reference electrode pair for $C_{1\text{-}9}$ is $C_{1\text{-}9}$, the reference electrode pair for $C_{1\text{-}10}$ is $C_{1\text{-}8}$, . . ., the reference electrode pair for $C_{1\text{-}16}$ is $C_{1\text{-}2}$. Table 1 enumerates all the measured electrode pairs and their corresponding reference electrode pairs for symmetrical cancellation when electrodes 1, 5, 9, 13 function as excitation electrodes. Electrode pairs corresponding to the remaining electrodes align with their reference electrode pairs, and so forth. Consequently, the reference data for symmetrical cancellation ECT originate from substance-field measurement data (Post-hemorrhage measurement), negating the need for measurements from a non-hemorrhaging cranium; thus, enabling the absolute imaging of cerebral hemorrhage. The imaging workflow for the symmetrical cancellation method is elucidated in Figure 2. Initially, the capacitance of all electrode pairs within a frame is measured, and the capacitance of all symmetrically offset reference electrode pairs is ascertained according to Table 1. Subsequently, the capacitance of each electrode pair in the original measurement frame is subtracted from the capacitance of the corresponding reference electrode pair, resulting in the ultimate imaging data for each electrode pair. Lastly, this finalized imaging capacitance data is integrated into the imaging algorithm to facilitate image reconstruction.

Imaging algorithm

The inverse problem within ECT seeks to reconstruct the permittivity distribution within an object based on capacitance measurements. In the case of minor change of permittivity $\Delta\varepsilon$, the relationship between the capacitance change ΔC and the change of permittivity $\Delta\varepsilon$ can be simplified to the following approximated linear equation:

$$\Delta C = S\Delta\varepsilon \tag{2}$$

Where, S is the sensitivity matrix, which is the prior information of image reconstruction that maps the permittivity distribution to capacitance change ΔC. Eq. 2 has to be discretized to calculate S and

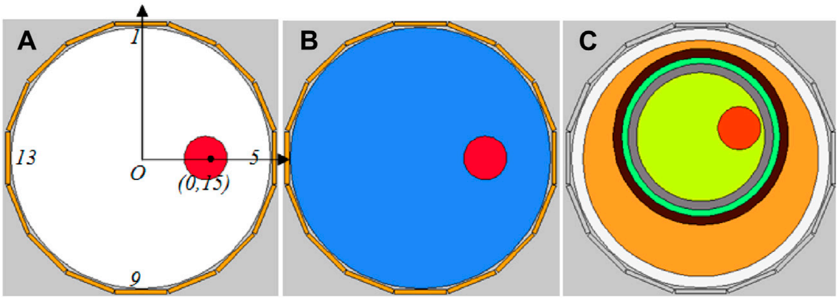


FIGURE 3
Simulation models of three cerebral hemorrhages. **(A)** Air-encapsulated cerebral hemorrhage model. **(B)** Water-encapsulated cerebral hemorrhage model. **(C)** Complex cerebral hemorrhage model comprising six distinct tissue.

TABLE 2 The relative permittivity of each part in the models A and B (1 MHz).

Model	A	B
Red (blood)	3000	3000
Background	1 (Air)	80 (Water)

visualize the permittivity distribution. The sensing area is divided into N elements or pixels. The discrete form of Eq. 2 can now be expressed as [21]:

$$\Delta C_{M\times 1} = S_{M\times N} \cdot g_{N\times 1} \tag{3}$$

where ΔC is the capacitance vector, g is the permittivity vector, i.e., the grey level of pixels in the imaging region, and S is the linearized sensitivity matrix, giving a sensitivity map for each electrode pair. M indicates the number of independent capacitance measurements in Eq. 1. The sensitivity map S is generally computed by the finite element simulation.

The sensitivity was calculated with the imaging zone under the air domain. The sensitivity of electrode pairs i-j at pixel point $P(x, y)$ is shown in Eq. 4, with (E_{xi}, E_{yi}) being the x-directional electric field component and the y-directional electric field component at pixel point P when electrode i is used as the excitation. (E_{xj}, E_{yj}) are the x-directional electric field component and the y-directional electric field component at pixel P when electrode j is used as the excitation. This air domain sensitivity matrix is used for both the simulation imaging and the later actual imaging.

$$\mathrm{S_{ij}}\left(\mathrm{x,y}\right) = -\mathrm{E_{xi}} \times \mathrm{E_{xj}} + \mathrm{E_{yi}} \times \mathrm{E_{yj}} \tag{4}$$

The inverse problem of ECT is to deduce the permittivity distribution $\varepsilon(\mathrm{x, y})$ from the measured capacitance vector ΔC. In its discrete form, the objective is to compute the unknown variable g from the known ΔC, employing Eq. 3, wherein S is considered a constant, *a priori* calculated matrix [22]. The resolution of this inverse problem constitutes the task of image reconstruction. Owing to the fact that the number of pixels N substantially exceeds the number of capacitance measurements M, Eq. 3 is ill-posed, rendering the solution non-unique. Therefore, reconstruction algorithms are imperative for the pursuit of an approximate solution.

In this paper, the Tikhonov regularization method is utilized to address the inverse problem of ECT [21]. The reconstructed distribution of permittivity g is ascertained as Eq. 5:

$$g = S^T\left(SS^T + \lambda I\right)^{-1}\Delta C \tag{5}$$

Where I is the identity regularization matrix, and λ is the regularization parameter which accounts for the degree of smoothness of the reconstructed image. The value was empirically selected and remained constant for the reconstructed images in subsequent sections. In order to overcome the issues of excessive smoothing and loss of information in the conventional Tikhonov regularization method, adaptive regularization methods should be considered in the future, where the regularization parameters are dynamically adjusted according to the characteristics of the data. Additionally, combining various regularization techniques such as Tikhonov regularization and L1 regularization (LASSO) could leverage the strengths of both to enhance the model's generalization ability and sparsity.

Simulation experiments

The simulation was executed using COMSOL Multiphysics and MATLAB on a computing environment equipped with an Intel Core i7 processor operating at 3.40 GHz. First and foremost, a 16-electrode ECT sensor model, depicted in Figure 3, was constructed in COMSOL, conforming to the dimensions of the actual ECT sensor employed in the subsequent physical experiment. Sixteen homogeneous rectangular electrodes (each possessing a width of 12 mm) were equidistantly positioned around a circle with a diameter of 60 mm. The internal surface of the 16 electrodes is encircled by a circle with a 60 mm diameter, constituting the imaging area. The 16 electrodes were enumerated in a clockwise orientation, with electrode 1 at the apex and electrode 9 at the nadir. For the resolution of the inverse problem, the imaging area is partitioned into a 32 × 32 grid, and the external portion of the circle is excluded, yielding 812 pixels within the imaging area. The sensitivity metrics for each of the 812 pixels were subsequently calculated for each electrode pair and utilized for imaging. Three distinct cerebral hemorrhage models (Figures 3A–C) were formulated for

TABLE 3 The relative permittivity of each part in the ICH model (1 MHz) [7].

Color	Orange	Black	Green	Gray	Yellow	Red
Typical tissues	Skin	Skull	Cerebrospinal fluid	Gray matter	White matter	Blood
Relative permittivity		150	108	991	700	3000

numerical simulation. The red circles in the triad of models serve to simulate hemorrhagic incidents, each with a diameter of 10 mm.

In both Models A and B, the coordinates of the red circle's center are (0 mm, 15 mm), strategically situated at the midpoint between the imaging area center *O* and electrode 5. The remainder of the imaging area in Models A and B, exclusive of the red circle, constitutes the background; Model A is denoted by white, and Model B is denoted by blue. Apart from the difference in the permittivity setting for the background, Models A and B are entirely same. The permittivity for each segment of Models A and B is specified in Table 2. The permittivity of the red circle is configured at 3000, equivalent to the permittivity of blood at 1 MHz. The background permittivity for Model A is set at 1, representing air, while the background for Model B is configured at 80, signifying water. Consequently, Model A represents an air-encapsulated cerebral hemorrhage model, and Model B represents a water-encapsulated cerebral hemorrhage model. Model C incorporates a complex cerebral hemorrhage model comprising six distinct tissue types. This model is grounded upon actual brain architecture, yet simplified by segmentation into six layers from the external to the internal, simulating skin (orange), skull (black), cerebrospinal fluid (green), gray matter (gray), white matter (yellow), and blood (red). The outermost layer is air. The relative permittivity of each component in Model C is calibrated to the measured values of human brain tissue, as documented in the literature [7] (Table 3). The small red circle denotes hemorrhaging in the right hemisphere, with central coordinates at (9 mm, 7 mm).

For each model, the data computed in accordance with the aforementioned permittivity parameters constitute the substance-field measurement data. This field measurement data is subtracted from the reference frame data to yield the final imaging data. Distinct reference data correspond to different imaging methodologies. For Model A, the reference data is the data obtained when the red sphere is eliminated and the imaging area is uniformly set to air. The substance-field data is subtracted from the reference data for imaging, signifying absolute imaging as the traditional manner. For Model B, traditional absolute imaging, time-differential imaging, and symmetric cancellation imaging are executed respectively. The reference data calculation scenarios for these three imaging methods are as follows: the imaging area is uniformly set to air (permittivity of 1), the imaging area is set to a blue background (permittivity of 80) post-elimination of the red circle, and the unaltered environment is congruent with substance-field measurement environment. For Model C, three modalities of imaging are also undertaken. The reference data calculation scenarios for these three imaging methodologies are as follows: the imaging area is uniformly set to air (permittivity of 1), the red circle is excised to retain other colored segments, and the unaltered environment is consistent with the substance-field measurement environment.

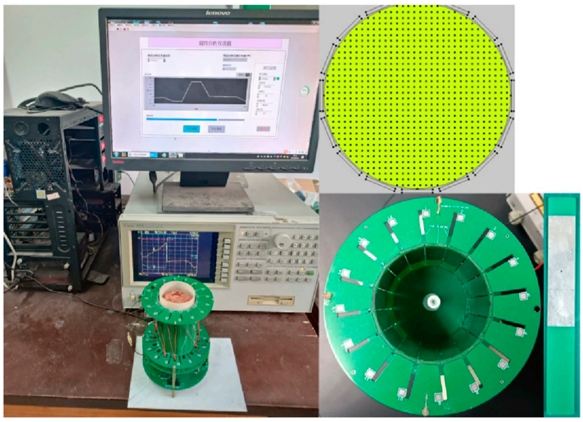


FIGURE 4
16-electrode ECT system.

To evaluate the quality of image reconstruction, the relative image error and the correlation coefficient between the true model and reconstructed images serve as assessment criteria. The definition of the relative image error and correlation coefficient is shown in Eqs 6, 7, respectively [22]. The lower the image error and the higher the correlation coefficient mean better image reconstruction outcomes.

$$Image\ error = \frac{\|\hat{g} - g\|}{\|g\|} \times 100\% \tag{6}$$

$$Correlation\ coe\,f\,ficient = \frac{\sum_{i=1}^{P}(g_i - \bar{g})(\hat{g}_i - \bar{\hat{g}})}{\sqrt{\sum_{i=1}^{P}(g_i - \bar{g})^2 \sum_{i=1}^{P}(\hat{g}_i - \bar{\hat{g}})^2}} \tag{7}$$

where $\hat{g}$ is the normalized pixel value reconstructed, and g is the normalized permittivity vector of a true distribution in the model. $\bar{\hat{g}}$, $\bar{g}$ respectively, are the mean values.

Physical model experiments

Subsequent to the simulation experiment, the 16-electrode ECT system designed in the preceding stage was utilized to image various physical models of cerebral hemorrhage employing diverse methodologies. The efficacy of various imaging methods was assessed to evaluate the feasibility of symmetric cancellation ECT imaging modalities. The 16-electrode ECT imaging system we employed is depicted in Figure 4. Its design originates from an impedance analyzer and is elaborated upon in the referenced literature [11].

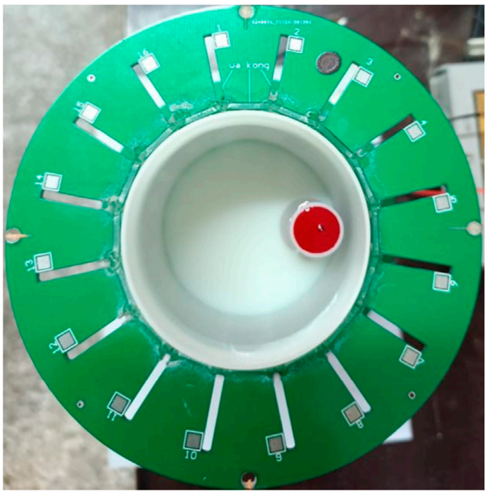


FIGURE 5
Water-wrapped blood hemorrhage model.

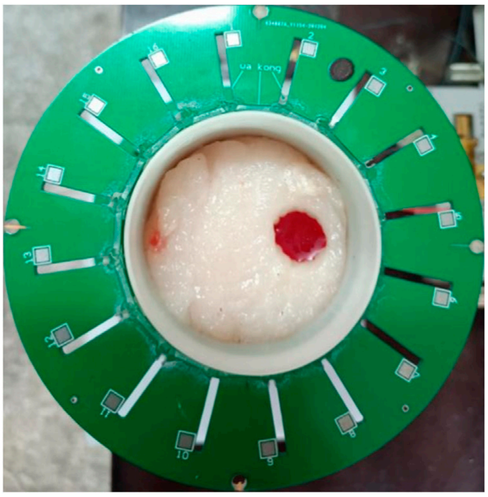


FIGURE 6
Fat-wrapped blood hemorrhage model.

The ECT Sensor comprises sixteen square electrodes, which are uniformly arranged on a circular base with a diameter of 60 mm. A singular electrode is fabricated from a square thin copper film (50 mm * 12 mm) imprinted on a PCB, incorporating a solder pad centrally located for welding electrode leads. The imaging area is a circle with a diameter of 60 mm, centered around the electrode circle and uniformly partitioned into 812 pixel points. The ECT Sensor can provide 16*15/2 = 120 independent capacitance measurements for the inverse calculating of the permittivities for the 812 pixels. An impedance analyzer (4294A, Agilent Technologies) was engaged to measure the capacitances of the electrode pairs. The excitation signal frequency is 1 MHz.

Imaging experiments of water-wrapped blood hemorrhage model

As shown in Figure 5, a 3D-printed cylinder (inner diameter 56 mm, outer diameter 58 mm) is equipped with a thin tube with a diameter of 10 mm. The center of the tube is situated 15 mm from the center of the cylinder. The interior of the tubule was inundated with fresh sheep blood following anticoagulation with heparin sodium, and the exterior of the tubule was inundated with distilled water. The elevation of the water surface is congruent with the elevation of the blood surface; both are 50 mm. Thus, in this model, blood serves as the imaging target, and water functions as the background. The cylinder injected with blood and distilled water is delicately positioned at the center of the ECT sensor and is coaxial with the imaging center. The thin tube filled with blood is aligned along the axis between the center of horizontal electrode 5 and electrode 13. The ECT system depicted in Figure 4 is utilized to measure a frame of data, constituting the substance-field measurement data. Subsequently, the blood in the thin tube is drained, and another frame of data is measured, serving as the reference frame data for the time-differential imaging. All the distilled water is then drained, and the entire interior of the barrel is rendered air-filled; another measurement ensues. The outcome is the reference data for traditional absolute imaging. The reference data for the symmetric cancellation method is extracted from the substance-field measurement data in accordance with the methodology outlined in Figure 2. Subsequently, the substance-field measurement data are subtracted from three disparate reference data for imaging, and the traditional absolute imaging, time-differential imaging, and symmetric cancellation imaging outputs are acquired respectively.

Imaging experiments of fat-wrapped blood hemorrhage model

As shown in Figure 6, a section of fresh porcine adipose tissue was procured from the market and reshaped using a cylindrical blade with an inner diameter of 56 mm. This yielded a cylindrical adipose specimen with a diameter of 56 mm and a height of 50 mm. Subsequently, another cylindrical blade, featuring an inner diameter of 10 mm, was utilized to excavate a cylindrical cavity with a diameter of 10 mm, situated 15 mm from the center of the adipose cylinder. This cavity was then inundated with the aforementioned sheep blood. The adipose cylinder engorged with blood was positioned at the center of the ECT Sensor's imaging area. Its orientation was adjusted to align the blood-filled cavity along the axis between the centroids of horizontal electrodes 5 and 13. In this model, blood serves as the imaging target while adipose tissue constitutes the background. Initially, the ECT system is engaged to capture a frame of data, constituting the substance-field measurement data. Thereafter, the blood within the adipose cavity is drained, followed by another data frame measurement, which serves as the reference data for time-differential imaging. The adipose samples are then removed, and the interior of the chamber is rendered air-filled; a subsequent measurement is taken. This yields the reference data for traditional absolute imaging. The reference data for the symmetric cancellation method is extracted from the substance-field measurement data. The substance-field measurement data are then subtracted from three distinct

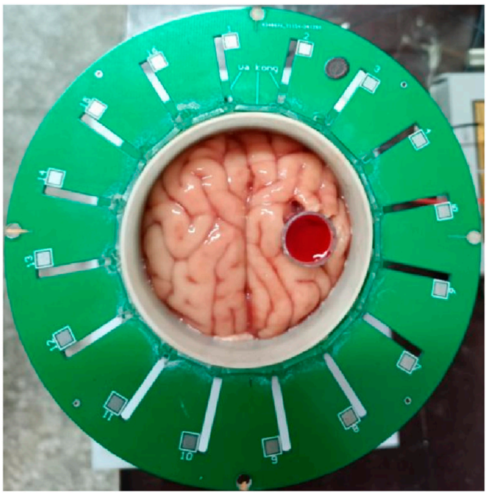


FIGURE 7
In vitro pig brain wrapped blood hemorrhage model.

reference data sets to generate imaging outputs: traditional absolute imaging, time-differential imaging, and symmetric cancellation imaging of the blood.

In vitro pig brain hemorrhage model imaging experiments

As shown in Figure 7, fresh porcine brain tissue is procured from the market and cautiously positioned within a cylinder featuring an inner diameter of 56 mm. It is gently compressed to preserve its anatomical structure, ensuring the symmetry of the left and right hemispheres about the longitudinal cerebral fissure. A syringe with a 10 mm diameter (lacking its tip) is inserted into the right hemisphere, and the syringe's center is situated approximately 13 mm from the center of the brain tissue. The syringe is then inundated with the aforementioned sheep blood. The cylinder, now filled with cerebral tissue and blood, is positioned at the center of the ECT Sensor's imaging area. By rotating the cylinder, the longitudinal fissure of the porcine brain is aligned with the axis connecting the centroids of electrodes 1 and 9, while the blood is located along the axis between the centroids of horizontal electrode 5 and 13. This orientation satisfies the requirement for symmetric cancellation imaging. In this model, blood is the imaging target and cerebral tissue serves as the background. Initially, the ECT system is engaged to capture a frame of data, which is the substance-field measurement data. Subsequently, the blood within the syringe is drained, followed by another data frame measurement, constituting the reference data for time-differential imaging. All cerebral tissue is then extracted, and the chamber is rendered air-filled; another measurement ensues. This produces the reference data for traditional absolute imaging. The reference data for the symmetric cancellation method is extracted from the substance-field measurement data. Thereafter, the substance-field measurement data are subtracted from three distinct reference data sets to yield imaging outputs for absolute imaging, time-differential imaging, and symmetric cancellation imaging of the blood.

Results and discussion

Simulation imaging results

The imaging results corresponding to the three simulation models in Figure 3 are delineated in Figure 8. The first line is the three original models of A, B and C, the second line is the traditional absolute imaging result corresponding to each model, the third line is the time-differential imaging result corresponding to each model, and the fourth line is the symmetrical cancellation imaging outcome corresponding to each model. For the absolute imaging utilizing the traditional method, only the result of Model A clearly delineates the location and dimension of the red blood circle, whereas the results of Model B and C negate the visualization of the blood spheres altogether. Owing to the fact that Model A is elementary and the background consists of air, its traditional absolute imaging is also time-differential imaging. The background of Model B and C is intricate. Particularly in Model C, the blood sphere is encased by quintuple layers of tissue. In this model, the capacitance change attributable to the blood sphere is entirely obscured by the capacitance change induced by the background, thereby corroborating that the traditional absolute imaging is unable to visualize the blood encased with a complex background. The time-differential imaging outcomes of the three models manifestly reflect the location and dimension of the blood spheres. However, the spatial localization of the image of blood of Model B and C exhibits a perceptible deviation relative to their positions in the original models, with both shifting towards the centroid. The white dotted circle in the imaging demarcates the original locale of the blood spheres in the model. This is predominantly attributable to the intricate backdrop. Because the sensitivity matrix S employed for imaging is calculated in the imaging region where the backdrop is a uniform permittivity distribution of air, and the background of Model B and C comprises heterogeneous distributions of varying permittivity. These distributions predominate over the expanse of the imaging area, and collectively converge towards the center of the imaging area; thus, the image of the blood in the imaging result shifts to the central point. Although time-differential imaging can visualize blood encased in intricate backdrops, it is inapplicable for the rapid detection of cerebral hemorrhage in clinical applications. This is predicated on the fact that the reference frame data requisite for time-differential imaging is the measurement devoid of hemorrhage, which can be simulated in computational environments, yet remains unattainable in clinical settings. The symmetrical cancellation imaging results of B and C explicitly demonstrate two symmetrical images of differing chromatic attributes, one pixel value being positive (red) and the other negative (blue). The chromatically red image is in precise correspondence with the position and dimension of the blood sphere in the model, indicating that the chromatically red image is indubitably the image of the blood in the model. Secondly, the pixel value is positive, indicative of an increase in the permittivity relative to the background, which is commensurate with the permittivity of actual blood. This congruence exists because the permittivity of the blood in the model exceeds that of the backdrop. It is justifiable to observe two symmetrical images with one positive and one negative pixel values in the symmetrical cancellation imaging results. This is corroborated by Figure 2, wherein the measurement capacitance of all electrode pairs in the symmetric cancellation mode is subtracted from the measurement capacitance of the electrode pairs symmetrically oriented about the intermediate symmetry axis. The capacitance of the

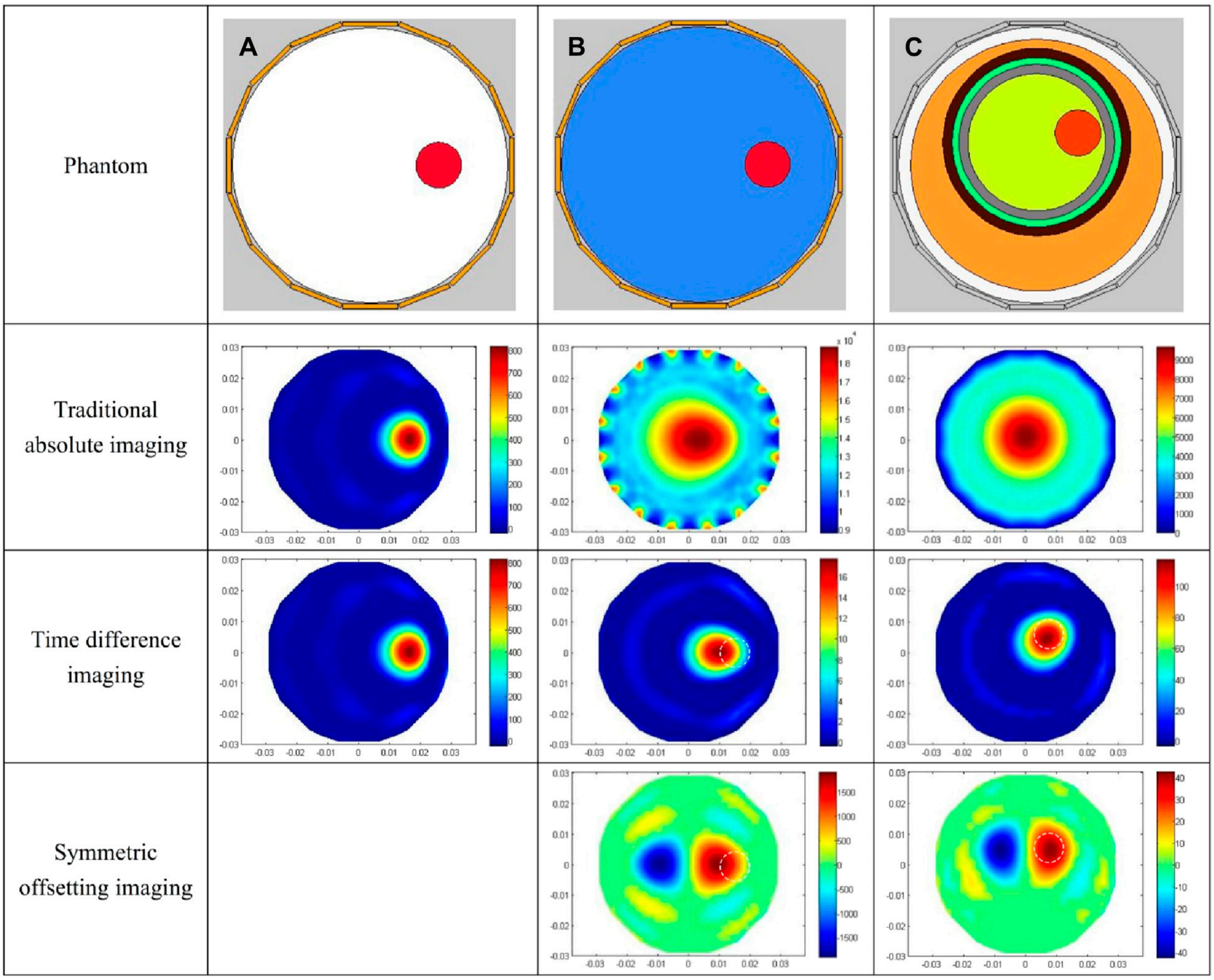


FIGURE 8
Simulation imaging results of three models. **(A)** Imaging results of the air-encapsulated cerebral hemorrhage model. **(B)** Imaging results of the water-encapsulated cerebral hemorrhage model. **(C)** Imaging results of the complex cerebral hemorrhage model comprising six distinct tissue.

electrode pairs of the hemorrhagic hemisphere is increased due to the presence of bleeding. In this context, the electrode pair capacitance of the hemorrhagic hemisphere subtracted from the electrode pair capacitance of the symmetrical non-hemorrhagic hemisphere constitutes a positive capacitance change. Conversely, the electrode pair capacitance of the same non-hemorrhagic hemisphere subtracted from the electrode pair capacitance of the symmetrical hemorrhagic hemisphere represents a negative capacitance change. The real hemorrhagic image in the outcomes of symmetric cancellation imaging can be prejudged based on the patient’s symptoms. The left and right neural centers of the brain severally govern the contralateral limb activity; thus, left cerebral hemorrhage predominantly induces right limb activity dysfunction, while right cerebral hemorrhage primarily provokes left limb activity dysfunction [23]. Additionally, the language center is localized in the left hemisphere, hence left cerebral hemorrhage can precipitate language dysfunction. The right brain orchestrates spatial imagination capabilities, and patients with right cerebral hemorrhage may manifest spatial imagination disorders. Moreover, experience can be accrued through imaging the cerebral hemorrhage ascertained by CT, thereby serving as a criterion for evaluating the real hemorrhage image in subsequent symmetrical cancellation imaging. A concomitant issue with symmetrical cancellation imaging results exists. The position of the blood image in BC model is likewise misaligned from its original locale in the model. In the figure, the white dotted circle demarcates the original position of the blood in the model. The reason for this result is the same as that of time-differential imaging. Model A did not undergo symmetrical cancellation imaging due to the simple background with air. The image error and correlation coefficient of different imaging results for the three models are shown in Table 4, and the symmetrical cancellation imaging results only consider the right red image.

Experimental results of physical model imaging

The results of the water-wrapped hemorrhage model are shown in Figure 9. The first row is the photographs of two original physical model. The blood-filled tubules in a horizontal orientation are proximal to electrode 13 (left) and electrode 5 (right),

TABLE 4 Image error (%) and Correlation coefficient for simulation results.

Model	Image error (%)			Correlation coefficient		
	A	B	C	A	B	C
Absolute imaging	9.58	86.56	94.27	0.96	0.23	0.18
Time-differential imaging	9.58	17.34	18.98	0.96	0.88	0.85
Symmetrical cancellation imaging		28.48	29.76		0.82	0.79

respectively. The center of the tubules is 15 mm away from the center of the imaging area. The second row displays the absolute imaging results with the traditional method. From the imaging results, the presence of blood cannot be seen at all. This predominantly stems from the fact that the volume of blood accounts for only 1/6 of the volume of the imaging area, which is much smaller than the volume of water. Subsequently, the permittivity of blood processed with heparin sodium is significantly attenuated compared to that of unadulterated blood, albeit marginally surpassing that of water [11]. These two reasons make it difficult for traditional absolute imaging to visualize water-wrapped blood. In the third row, two time-differential imaging images can clearly show the existence of blood, and the position and size of the circular image are basically the same as the position and size of the blood in the actual model, but its position is shifted to the center of the imaging area by about 5 mm. The white dotted circle in the image indicates the position of the actual blood. This phenomenon corroborates the outcome of the time-differential imaging in the simulation of Figure 8, and the reason is the same. The fourth row exhibits the symmetrical cancellation imaging results. Each image clearly shows two circular images of the same shape and size, one red and one blue, and is symmetrical about the axis of symmetry of the ECT Sensor. By analyzing the imaging data, it is found that the pixel values of the two symmetrical images are also complementary; ergo, one is positive, and one is negative. The pixel value of the red image is positive, and the value of the blue image is negative. Since the permittivity of the blood exceeds that of the ambient water, the red circular image represents the image of the actual blood, which also corresponds to the position of the blood in the actual model. Thus, symmetric cancellation imaging can indeed execute absolute imaging of water-wrapped blood without the need for reference data devoid of bleeding, which is not possible with time-differential imaging. Nonetheless, the noise in the symmetrical cancellation imaging result is much larger than the time-differential imaging. In addition to the two blood images of one red and one blue, numerous artifacts reside on the edge. This is predominantly attributable to the position deviation of the bucket in the actual model. Since the external diameter of the bucket is 58 mm and the diameter of the imaging area is 60 mm, there exists a gap of 1 mm between the edge of the bucket and the electrode surface. In practice, it is difficult to guarantee that the center of the bucket completely coincides with the center of the imaging area, thereby engendering a variable gap size between the upper and lower and left and right boundaries of the bucket and the corresponding electrodes. This inconsistency culminates in left and right asymmetry, and the accuracy of symmetry cancellation imaging entirely depends upon the symmetry of the upper and lower and left and right sectors. The higher the symmetry, the better the imaging quality. The asymmetry of the left and right sides of the bucket will engender symmetrical image noise on the left and right edges, and the asymmetry of the upper and lower sides of the bucket will also cause symmetrical image noise on the upper and lower edges. Because the overall volume of the barrel is much larger than the blood, very small asymmetry will cause a large image noise. The position of the blood image in the symmetrical cancellation imaging is also approximately 5 mm away from the center relative to its actual position, which is the same as the result in the simulation, and the reason is the same. The image error and correlation coefficient for the three imaging results of the two models are shown in Table 5. The image error of the time-differential imaging is the smallest, the correlation coefficient is the best, and the second is the symmetrical cancellation imaging. Because the result of absolute imaging is too poor, two imaging quality metrics are not included.

The results of the fat-wrapped hemorrhage model are shown in Figure 10. In the two prototypes in the first row, the blood in the horizontal cylindrical cavity is proximate to electrode 13 (left) and electrode 5 (right), respectively, and the center of the blood is 15 mm away from the center of the imaging area. The absolute imaging results of the second row are completely red, and the existence of blood cannot be seen at all, which proves that the absolute imaging with the traditional method cannot image the blood wrapped by fat tissue. The time-differential imaging results of the third row can clearly show the image of the blood. As in the scenario of water-wrapped hemorrhage model, the position of the blood image is shifted to the center by several millimeters relative to its position in the actual model. The white dotted circle in the figure demarcates the actual locus of the blood. In the symmetrical cancellation imaging results of the final row, both images clearly show two images with the same shape and size, one red and one blue color, and is symmetrical about the axis of the ECT Sensor. As in the case of the water-wrapped hemorrhage model, the pixel values of the two symmetrical images are equivalent, albeit with antithetical signs. As the permittivity of the blood is also greater than that of fat, so the positive red image is the actual image of the blood, which also corresponds to the position of the blood. Similarly, the noise in the symmetrical cancellation imaging result is much larger than that of the time-differential imaging, and many small patches appear on the edge. This is mainly due to the inconsistent gap size between the outside surface of the barrel and the electrode array, which is caused by poor symmetry. Secondarily, owing to limited manufacturing precision, the fat structure lacks a standardized cylindrical configuration, and its surface is not entirely planar, which further compromises bilateral symmetry. The position of the blood image deviates by several millimeters toward the center relative to its actual localization. In comparison to the symmetrical cancellation imaging results of the water-wrapped blood model in Figure 9, it is found that the dual symmetrical images of the blood in Figure 10, irrespective of hue,

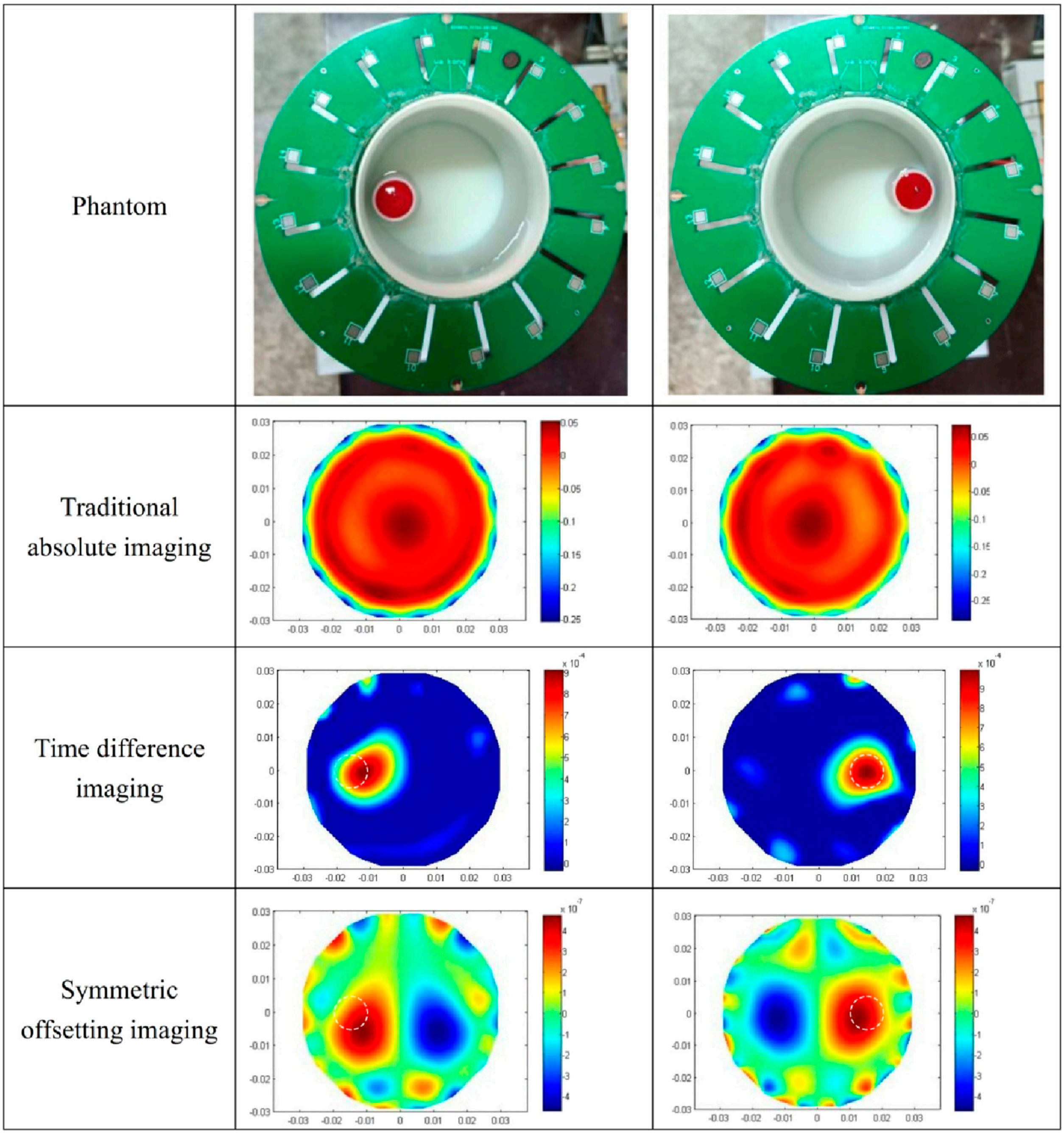


FIGURE 9
The results of three kinds of imaging methods for two water-wrapped blood models.

TABLE 5 Image error (%) and Correlation coefficient for imaging results of two water-wrapped blood models.

	Image error (%)		Correlation coefficient	
	Blood on the left	Blood on the right	Blood on the left	Blood on the right
Time-differential imaging	25.47	26.52	0.78	0.77
Symmetrical cancellation imaging	35.35	35.69	0.73	0.71

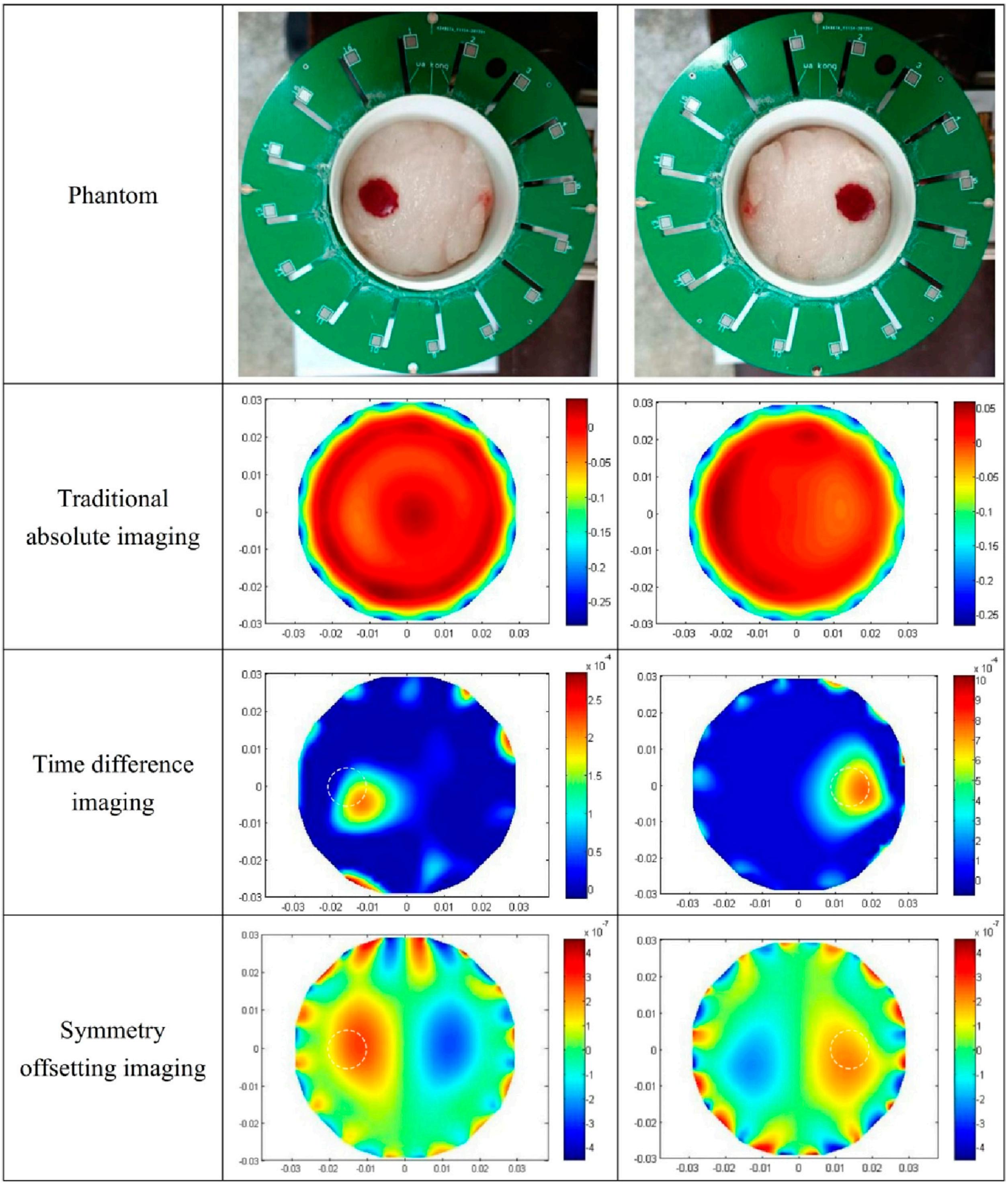


FIGURE 10
The results of three kinds of imaging methods for two fat-wrapped blood models.

are diminished in intensity. That is, the contrast differential between the pixel values of the blood image and the background is attenuated. This is mainly because the permittivity difference between fat and blood is inferior to that between water and blood, rendering it more challenging to visualize fat-wrapped blood. The image error and correlation coefficient for the three imaging modalities are shown in Table 6. The image error of time-differential imaging is the smallest, the correlation coefficient is the best, followed by symmetrical cancellation imaging. Due to the poor performance in absolute imaging, two quantitative metrics are not included.

The imaging results of the isolated porcine brain wrapped hemorrhage model are shown in Figure 11. In the two prototypes

TABLE 6 Image error (%) and Correlation coefficient for imaging results of two fat-wrapped blood models.

	Image error (%)		Correlation coefficient	
	Blood on the left	Blood on the right	Blood on the left	Blood on the right
Time-differential imaging	27.55	25.64	0.73	0.75
Symmetrical cancellation imaging	37.65	38.76	0.71	0.70

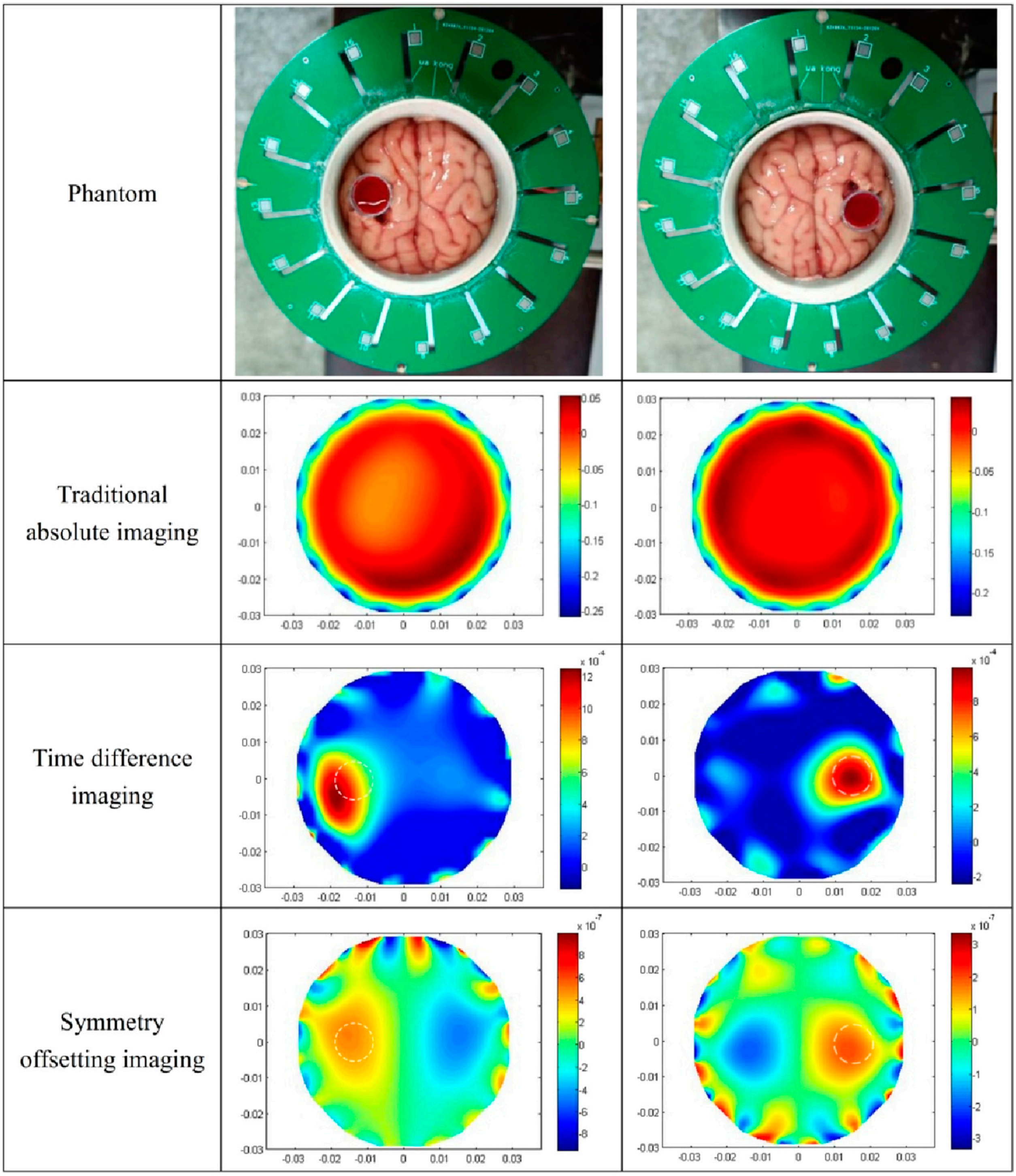


FIGURE 11
The results of three kinds of imaging methods for two isolated porcine brain wrapped blood models.

TABLE 7 Image error (%) and Correlation coefficient for imaging results of two porcine brain wrapped hemorrhage models.

	Image error (%)		Correlation coefficient	
	Blood on the left	Blood on the right	Blood on the left	Blood on the right
Time-differential imaging	29.12	28.38	0.71	0.72
Symmetrical cancellation imaging	39.44	38.25	0.69	0.68

presented in the first row, the syringes filled with blood in a horizontal orientation are proximate to electrode 13 (left) and electrode 5 (right), respectively, and the center of the syringe is 13 mm from the center of the imaging area. The absolute imaging results with the traditional method of the second row are wholly suffused in red hue, rendering the presence of blood indiscernible. This is the same as the absolute imaging results of the previous two models, which proves that the traditional absolute imaging cannot image the blood wrapped in the porcine brain. The time-differential imaging results of the third line can clearly show the images of the blood in the two models, and the position of the blood image is shifted to the center by several millimeters relative to the position of the blood in the actual model. The white dotted circle in the figure represents the actual locus of the blood. In the symmetrical cancellation imaging results of the final row, both images clearly show two images with the same shape and size, one in a red hue and another in blue, symmetrical about the axis of the ECT Sensor. Consistent with the preceding models, the pixel values of the two symmetrical images are equivalent, albeit with antithetical signs. As the permittivity of the blood is also greater than that of other brain tissues, so the positive red image is the actual image of the blood, which also corresponds to the position of the blood in the actual model. Similarly, the noise in the symmetrical cancellation imaging result is much larger than that in the time-differential imaging, and numerous noises manifest at the periphery of the imaging area. This predominantly emanates from the inconsistent gap size between the outside surface of the barrel and all electrodes due to the asymmetry. Secondarily, due to the inherent pliability of the porcine brain tissue and manual stacking, the resultant structure lacks a standardized cylindrical configuration and bilateral symmetry is compromised. Furthermore, in an attempt to preserve the original morphology of the left and right hemispheres, the surface remains non-planar, thus impacting the symmetry. The size and position of the blood image in the two models in the final row vary, attributed to the relative position discrepancy of the barrel in the two models. The spatial orientation of the blood image deviates minimally toward the center, yet this deviation is less than in previous models. This is attributed to the nearer proximity of the actual syringe center, being only 13 mm from the imaging area's center. The reason behind this procedural adjustment lies in the intrinsic softness of the porcine brain tissue; positioning the syringe closer to the edge would destroy its overall structural integrity. Analogous to the fat-wrapped blood scenario, the contrast differential between the pixel values of the blood image and the background image is attenuated in Figure 11. This is mainly because the permittivity difference between the porcine brain tissue and blood is inferior to that between water and blood, making it more difficult to visualize porcine brain tissue-wrapped blood. The image error and correlation coefficient of imaging results are shown in Table 7. Similarly, the image error of time-differential imaging is the smallest, the correlation coefficient is the best, followed by symmetrical cancellation imaging. Due to the poor performance in absolute imaging, two quantitative metrics are not included.

Conclusion

At present, ECT can only image cerebral hemorrhage with time-differential imaging. Time-differential imaging requires measurement data when the patient is not bleeding, but this is difficult to achieve in practice. Therefore, this imaging modality lacks the capability for rapid acquisition of the absolute image information of cerebral hemorrhage, thus rendering it unable for rapid diagnostic applications. To solve this limitation, a symmetrical cancellation ECT Imaging method was proposed, predicated upon the anatomical symmetry between the left and right cerebral hemispheres. This method only needs that the sagittal suture of the examined cranium remain collinear with the central axis (symmetry axis) of a corresponding pair of electrodes in the ECT sensor. Such alignment enables the electrodes on both sides to maintain symmetry about the sagittal suture. Consequently, imaging data is attainable through the subtraction of capacitance measurements from their symmetrical counterparts. The reference data for this novel imaging modality is directly derived from the measurement data after bleeding, thereby eliminating the need for pre-hemorrhage measurement. In order to verify the feasibility of this scheme, simulation and empirical imaging evaluations were conducted across various cerebral hemorrhage models. The findings corroborate that the imaging method can indeed facilitate the absolute imaging of cerebral hemorrhage in the established way. Moreover, the results of this imaging method have a significant feature, that is, an artifact with the same size and shape and the opposite pixel value symbol appears on the opposite side of the actual bleeding image. This is determined by the principle of symmetrical cancellation imaging, no need to worry. Clinically, the hemisphere in which hemorrhage exists can be easily determined through patient's symptoms. Moreover, the imaging quality is intrinsically dependent upon the anatomical symmetry across the cerebral hemispheres around the electrode's symmetry axis; superior symmetry yields enhanced imaging quality and diminished noises. Nonetheless, practical applications are encumbered by operational complexities and the imperative for precise cranial alignment. The blood locus within the symmetrical cancellation imaging exhibits a slight deviation from its true coordinates, primarily due to non-homogeneous distribution of the permittivity within the imaging domain. Brain belongs to the non-uniform dielectric distribution, which contains tissues such as gray matter white matter cerebrospinal fluid, and also contains a small amount of residual blood, so it is a very inhomogeneous medium, thus leading to a large difference between the sensitivity distribution of the imaging area full of brain and that full of air, and the sensitivity matrices used for imaging in this paper are all calculated when the imaging zone are full of air, so it leads to poor imaging of blood, which may also be the reason why the imaging results deviate from the actual location. This problem can be potentially mitigated through refinement in the computational approaches for sensitivity matrix and imaging algorithms. In summation, the Symmetrical Cancellation Imaging modality elucidated here demonstrates potential for achieving absolute cerebral hemorrhage imaging, although further research is require for entering practical stage. The most significant hurdles pertain to the intricacy of operational procedures and the precision required in cranial placement. Future directions involve the refinement of ECT sensor design, computational methodologies for sensitivity matrix and imaging algorithms to enhance the imaging quality for this innovative imaging paradigm. To ensure that the ECT electrode array maintains precise symmetry with respect to the skull's anatomical structure, it is proposed to install laser range sensors in the center of each electrode. This setup will allow for real-time display of the distances between all electrodes and the subject's skull. Consequently, based on the measurements from the laser range sensors, the gap between the ECT electrode array and the skull can be adjusted more conveniently, significantly improving the precision of the array's symmetry relative to the skull. Additionally, procedure for the calibration of symmetrical cancellation ECT is needed. It is proposed to design a 3D-printed adult skull model based on 3D scanning data of an adult skull. The model will include major fillers such as cerebrospinal fluid, gray matter, and white matter, whose electromagnetic parameters match those of their real-life counterparts. Before each measurement, the model will be measured first by the symmetrical cancellation ECT, the measurement data will be stored as calibration data. Subtracting this calibration data from subsequent actual measurements can achieve high-precision imaging results.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

JH: Data curation, Writing–original draft. FC: Data curation, Writing–original draft. KW: Conceptualization, Supervision, Writing–review and editing. SC: Conceptualization, Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol* (2019) 18:439–58. doi:10.1016/S1474-4422(19)30034-1

2. Wang L, Liu J, Yang G, Peng B, Wang Y. The prevention and treatment of stroke in China is still facing great challenges—summary of China Stroke Prevention report 2018. *Chin Circ J* (2019) 34(02):6–20.

3. Ayaz H, Lzzetoglu M, Izzetoglu K, Onaral B, Ben Dor B. Early diagnosis of traumatic intracranial hematomas. *J Biomed Opt* (2019) 24(5):1. doi:10.1117/1.jbo.24.5.051411

4. Balami JS, White PM, McMeekin PJ, Ford GA, Buchan AM. Complications of endovascular treatment for acute ischemic stroke: prevention and management. *Int J Stroke* (2018) 13(4):348–61. doi:10.1177/1747493017743051

5. Bayford R, Bertemes-Filho P, Frerichs I. Topical issues in electrical impedance tomography and bioimpedance application research. *Physiol Meas* (2020) 41(12): 120301. doi:10.1088/1361-6579/abcb5b

6. Soleimani M, Ktistis C, Ma X *Magnetic induction tomography: image reconstruction on experimental data from various applications* (2005).

7. Sasaki K, Wake K, Watanabe S. Development of best fit Cole-Cole parameters for measurement data from biological tissues and organs between 1 MHz and 20 GHz. *Radio ence* (2015) 49(7):459–72. doi:10.1002/2013RS005345

8. Warsito W, Marashdeh Q, Fan LS. Electrical capacitance volume tomography. *IEEE Sensors J* (2007) 7(4):525–35. doi:10.1109/jsen.2007.891952

9. Wang H, Yang W. Application of electrical capacitance tomography in pharmaceutical fluidised beds – a review. *Chem Eng Sci* (2020) 231:116236. doi:10.1016/j.ces.2020.116236

10. Bai Z, Li H, Chen J, Zhuang W, Li G, Chen M, et al. Research on the measurement of intracranial hemorrhage in rabbits by a parallel-plate capacitor. *PeerJ* (2021) 9(99): e10583. doi:10.7717/peerj.10583

11. Xu R, Zhuang W, Bai Z, Wang F, Jin G, Liu N, et al. A pilot study on intracerebral hemorrhage imaging based on electrical capacitance tomography. *Front Phys* (2023) 3(11). doi:10.3389/fphy.2023.1165727

12. Chen Q, Liu R, Wang C, Liu RJMS. Real-time *in vivo* magnetic induction tomography in rabbits: a feasibility study. *Meas Sci Technol* (2020) 32(3):035402. doi:10.1088/1361-6501/abc579

13. Roldan-Valadez E, Suarez-May MA, Favila R, Aguilar-Castañeda E, Rios C. Selected gray matter volumes and gender but not basal ganglia nor cerebellum gyri discriminate left versus right cerebral hemispheres: multivariate analyses in human brains at 3T. *anatomical Rec Adv Integr Anat Evol Biol* (2015) 298(7):1336–46. doi:10.1002/ar.23165

14. Samir A. Anatomy, histology of cerebral hemisphere then the classification of CNS tumors with referral to morphology, complications and prognosis of Astrocytoma. *Meningioma* (2021). doi:10.13140/RG.2.2.14518.52802

15. Gm DCM, Brott TG, Xi G *Lobar intracerebral hemorrhage model in pigs: rapid edema development in perihematomal white matter* (2013).

16. Wang J, Li C, Chen T, Fang Y, Shi X, Pang T, et al. Nafamostat mesilate protects against acute cerebral ischemia via blood–brain barrier protection. *Neuropharmacology* (2016) 105:398–410. doi:10.1016/j.neuropharm.2016.02.002

17. Dai M, Liu XC, Li HT, Xu CH, Yang B, Wang H, et al. EIT imaging of intracranial hemorrhage in rabbit models is influenced by the intactness of cranium. *Biomed Res Int* (2018) 2018:1–10. doi:10.1155/2018/1321862

18. Jin G, Sun J, Qin M, Tang Q, Xu L, Ning X, et al. A new method for detecting cerebral hemorrhage in rabbits by magnetic inductive phase shift. *Biosens Bioelectron* (2014) 52:374–8. doi:10.1016/j.bios.2013.09.019

19. Mcdermott BJ, Porter E, Jones M, McGinley B, O'Halloran M. Symmetry difference electrical impedance tomography - a novel modality for anomaly detection. *Physiol Meas* (2018) 39:044007. doi:10.1088/1361-6579/aab656

20. Mcdermott B, Avery J, O'Halloran M, Aristovich K, Porter E. Bi-frequency symmetry difference electrical impedance tomography—a novel technique for perturbation detection in static scenes. *Physiol Meas* (2019) 40:044005. doi:10.1088/1361-6579/ab08ba

21. Yang WQ, Peng L. Image reconstruction algorithms for electrical capacitance tomography. *Meas Sci Tech* (2003) 14:R1–R13. doi:10.1088/0957-0233/14/1/201

22. Ye J, Wang H, Yang W. Image reconstruction for electrical capacitance tomography based on sparse representation. *IEEE Trans Instrumentation Meas* (2014) 64(1):89–102.

23. Gao J, Yang C, Li Q, Chen L, Jiang Y, Liu S, et al. Hemispheric difference of regional brain function exists in patients with acute stroke in different cerebral hemispheres: a resting-state fMRI study. *Front Aging Neurosci* (2020) 13:691518. doi:10.21203/rs.3.rs-126624/v1

TYPE Original Research
PUBLISHED 04 June 2024
DOI 10.3389/fphy.2024.1387717

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Mianyi Chen,
Xverse Co., Ltd., China
Yixin Liu,
China Academy of Engineering Physics, China
Yongfeng Liu,
Yangzhou University, China
Yukun Huang,
Shanghai United Imaging Medical Technology Co., Ltd., China
Maria Del Rocio Camacho Morales,
Australian National University, Australia

*CORRESPONDENCE
Min Zhao,
✉ zhaomin@cqu.edu.cn

RECEIVED 18 February 2024
ACCEPTED 06 May 2024
PUBLISHED 04 June 2024

CITATION
Wang B, Long Z, Chen X, Feng C, Zhao M, Sun D, Wang W and Wang S (2024), Research on LiDAR point cloud data transformation method based on weighted altitude difference map.
Front. Phys. 12:1387717.
doi: 10.3389/fphy.2024.1387717

Research on LiDAR point cloud data transformation method based on weighted altitude difference map

Bosi Wang[1,2,3], Zourong Long[4], Xinhai Chen[1,2,3], Chenjun Feng[3], Min Zhao[1]*, Dihua Sun[1], Weiping Wang[5] and Shihao Wang[4]

[1]College of Automation, Chongqing University, Chongqing, China, [2]China Merchants Auto-trans Technology Co., Ltd., Chongqing, China, [3]China Merchants Testing Vehicle Technology Research Institute Co., Ltd., Chongqing, China, [4]Chongqing University of Technology, Chongqing, China, [5]Chongqing Expressway Group Co., Ltd., Chongqing, China

Road surface detection plays a pivotal role in the realm of autonomous vehicle navigation. Contemporary methodologies primarily leverage LiDAR for acquiring three-dimensional data and utilize imagery for chromatic information. However, these approaches encounter significant integration challenges, particularly due to the inherently unstructured nature of 3D point clouds. Addressing this, our novel algorithm, specifically tailored for predicting drivable areas, synergistically combines LiDAR point clouds with bidimensional imagery. Initially, it constructs an altitude discrepancy map via LiDAR, capitalizing on the height uniformity characteristic of planar road surfaces. Subsequently, we introduce an innovative and more efficacious attention mechanism, streamlined for image feature extraction. This mechanism employs adaptive weighting coefficients for the amalgamation of the altitude disparity imagery and two-dimensional image features, thereby facilitating road area delineation within a semantic segmentation framework. Empirical evaluations conducted using the KITTI dataset underscore our methodology's superior road surface discernment and extraction precision, substantiating the efficacy of our proposed network architecture and data processing paradigms. This research endeavor seeks to propel the advancement of three-dimensional perception technology in the autonomous driving domain.

KEYWORDS

road vehicles, convolutional neural nets, image processing, data fusion, semantic segmentation

1 Introduction

In the evolving landscape of intelligent transportation, the escalating demand for precision in perception algorithms renders single image sensor modalities inadequate. Visual imagery is susceptible to ambient light intensity variations, where shadows cast by tall structures and trees can precipitate algorithmic inaccuracies or omissions. In scenarios devoid of depth information, conventional visual image-based algorithms exhibit limited efficacy in discerning road edges and pedestrian crossings. Conversely, LiDAR radar, impervious to lighting and shadows, provides high-precision environmental depth data, enhancing detection stability significantly. Perceiving road information using LiDAR point cloud, which is collected by LiDAR sensors, is both a challenging research area and a key focus in the field.

Several researchers have explored LiDAR-based road information extraction techniques. Zhang et al. [1] utilized Gaussian difference filtering for point cloud segmentation, aligning the results with a model to isolate ground points. Chen et al. [2], targeting lane edge information, segmented lanes post feature extraction. Asvadi et al. [3] adopted segmented plane fitting as their evaluative criterion. Wijesoma et al. [4] approached the challenge by focusing on road edge detection, employing extended Kalman filtering for lane edge feature extraction.

The fusion of LiDAR and camera data for road perception has garnered increasing scholarly interest. The inherent disparity between three-dimensional LiDAR point clouds and two-dimensional image pixels presents a significant data space challenge. Innovative algorithms have been developed to transform and densify sparse point cloud data into continuous, image-like formats. Chen et al. [5] leveraged LiDAR's scanning angle data to create image-like representations from point clouds. Thrun et al. [6] introduced a top-down radar feature representation based on vertical point cloud distribution. Gu et al. [7] employed linear upsampling for point cloud data preprocessing, extracting features from the densified clouds for road perception. Similarly, Fernandes et al. [8] utilized upsampling but projected the point cloud onto the X-Y plane before extracting *Z*-axis height values. Caltagirone et al. [9] generated a top view of point clouds by encoding their average degree and density, facilitating road perception. Han X et al. [10] and Liu Z et al. [11] further contributed with high-resolution depth image generation and directional ray map implementation, respectively.

Existing methods that densify point clouds into more manageable data forms often lead to computationally intensive outputs, compromising the real-time capabilities of the overall algorithm. To address this, our paper introduces a novel method for 3D point cloud conversion, leveraging weighted altitude differences. This approach not only efficiently preserves essential road information but also enhances the distinction between road and non-road areas.

In this study, we propose distinct fusion strategies at both the data and feature levels, tackling the challenges posed by disparate sensor data structures and varied road characteristics. Initially, we transform three-dimensional point cloud data into a two-dimensional weighted altitude difference map. This process, anchored on the uniform height variation in flat road areas, not only retains crucial road features but also facilitates data-level fusion. Subsequently, we introduce a LiDAR-camera feature adaptive fusion technique. This innovative method refines the semantic segmentation network encoder and integrates a feature adaptive fusion module. This module, comprising an adaptive feature transformation network and a multi-channel feature weighting cascade network, adeptly linearly transforms LiDAR radar features. These transformed features are then coalesced with visual image features across multiple levels, achieving effective feature-level fusion of multimodal data.

2 Weighted altitude difference map based on point cloud data

2.1 Altitude difference map

The disparity between original LiDAR data and visual data presents significant challenges in direct data fusion and feature extraction. LiDAR data, comprising tens of thousands of points in a three-dimensional space, assigns each point with 3D coordinates (x, y, z). In contrast, visual data consists of an array of pixels on a two-dimensional image plane, each pixel defined by an RGB value. This fundamental difference in data space complicates their direct integration.

In the context of road areas, the LiDAR point cloud exhibits a unique smoothness compared to other objects. This smoothness is evident as the road area's point cloud in 3D space shows fewer irregularities, unlike non-road areas and entities like vehicles and pedestrians. The discontinuities in the point cloud bounding box are more pronounced for these non-road elements. The road surface's smoothness is quantified by the minimal average altitude difference between road surface points and their neighboring points.

Through the process of joint calibration parameters and sparse point cloud densification, a detailed projection image of the dense LiDAR point cloud is obtained. This involves projecting the 3D coordinate vectors of the LiDAR points onto a 2D image plane, resulting in varying shapes depending on the observation coordinates along the X, Y, and *Z*-axes. By defining the X-Y plane as the base, the *Z*-axis can be interpreted as the height value of the point cloud, providing a crucial dimensional perspective.

As shown in Figure 1A, the absolute value of the altitude difference between two positions (such as Z_0 and Z_i in the Figure 1) is calculated as the spatial displacement between them. The specific formula for the altitude difference value $g_{x,y}$ located at (x_0, y_0) is as follows:

$$g_{x,y} = \frac{1}{M} \sum_i |Z_i - Z_0|$$

In the formula, Z_0 represents the height on the *Z*-axis of the point projected at the coordinate (x_0, y_0), Z_i represents the height on the *Z*-axis of other points in the neighborhood of point (x_0, y_0), and M represents the total number of points to be considered in the set neighborhood.

Finally, all calculated $g_{x,y}$ values are scaled between 0–255, and the scaled $g_{x,y}$ is used as the gray value at point (x, y) on the image to form a gray-scale image with the altitude difference value as the pixel value. This can be regarded as a two-dimensional image plane composed of the average altitude difference values of the projected points. The resulting altitude difference gray-scale image is shown in Figure 1B.

The relationship between the average altitude difference of a point relative to its neighbors and the resultant grayscale value in the converted height map is inversely proportional. As illustrated in Figure 1B, an upright and sharply defined object will cast a projection with a significant altitude difference on the image plane. Consequently, the road area, characterized by minimal intensity, appears darker in the image. In contrast, other objects typically exhibit higher altitude values, resulting in more pronounced intensity differences when compared to the road area. This conversion from original 3D data to point cloud altitude difference effectively encapsulates the road's inherent characteristics and smoothness present in the initial LiDAR data. The height map thus produced simplifies the task for a deep convolutional neural network model in discerning and

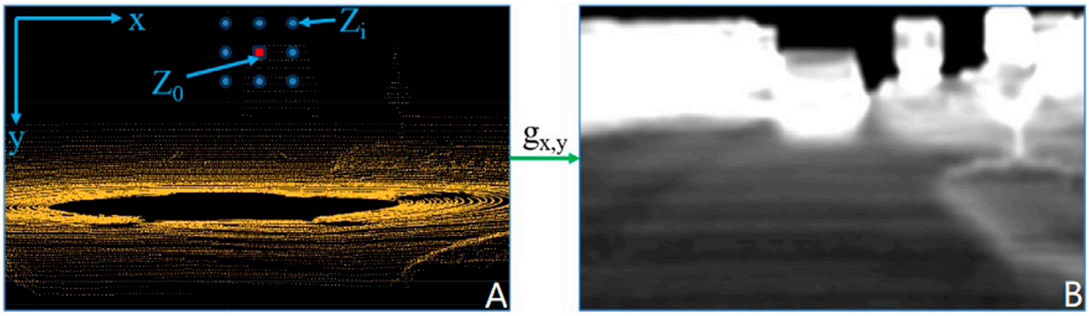


FIGURE 1
Altitude difference image conversion process. **(A)** The point cloud image, **(B)** the calculated altitude difference image.

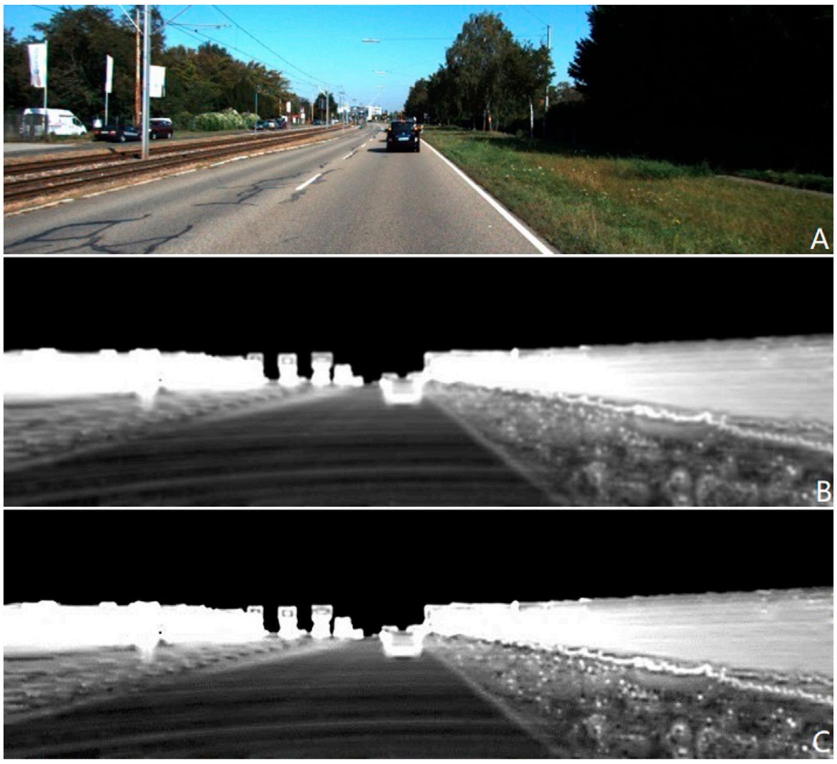


FIGURE 2
Point cloud data conversion results, **(A)** is the RGB image, **(B)** is the original Altitude Difference Map, and **(C)** is the Weighted Altitude Difference Map. **(C)** Contains more details, and the changes in height are more pronounced in the pixel values.

identifying the road, enhancing the model's ability to differentiate between various features.

2.2 Weighted altitude difference map

The elevation difference image principally focuses on the height variation between a central point and its surrounding points. Upon examination, it becomes apparent that the low grayscale values in road areas on this image stem from the negligible height changes extending in all directions from any given point on the road, leading to minimal elevation difference values. Conversely, the areas of higher intensity on the elevation difference image are predominantly located where road and non-road areas intersect. These high-intensity regions usually align approximately along the Y-axis. A marked change in elevation difference values is observed when neighboring points along the X-axis direction are selected for calculation, distinguishing them from the road surface area.

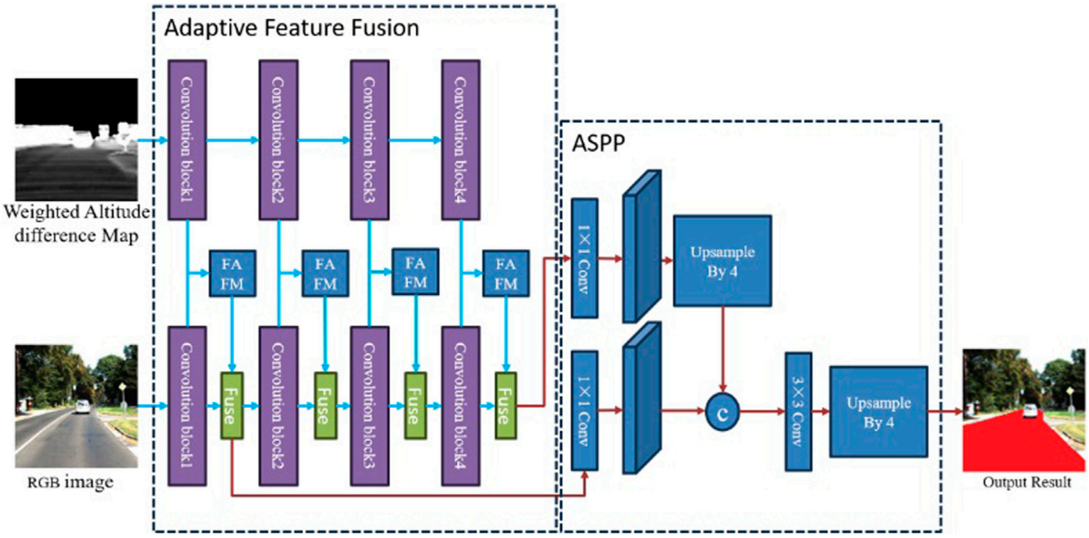


FIGURE 3
Feature adaptive fusion network.

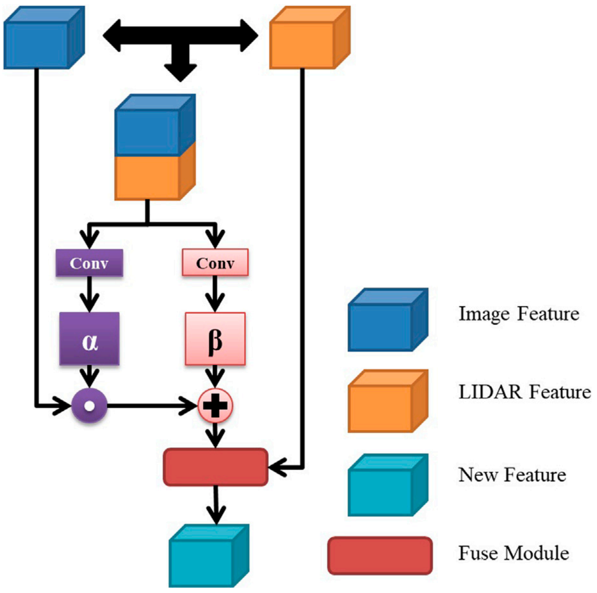


FIGURE 4
FAFM.

To leverage this characteristic, we propose an enhanced elevation difference conversion method. The novel formula for calculating elevation difference values is structured to more accurately reflect these spatial variations. This approach aims to provide a clearer distinction between road and non-road areas, improving the precision of the elevation difference image for subsequent analysis and application. The new formula for calculating elevation difference values is as follows:

$$g_{x,y} = \max\left(\frac{1}{M}\sum_i \gamma_{1i}\cdot|Z_i - Z_0|, \frac{1}{M}\sum_i \gamma_{2i}\cdot|Z_i - Z_0|\right)$$

$$\gamma_{1i} = Sigmoid\,(X_i - X_0) + 0.5$$
$$\gamma_{2i} = Sigmoid\,(X_0 - X_i) + 0.5$$

In the formula, X_0 and X_0 respectively represent the X-axis coordinates of the center point and the neighborhood point, and γ_{1i} and γ_{2i} are adaptive weight parameters. When the center point is located in the road surface area, the introduction of new weight calculation will not cause an increase in numerical intensity. When the center point is located near the left or right boundary, the characteristic of the drastic increase in elevation difference will be amplified by one of the adaptive weight parameters γ_{1i} and γ_{2i}. The amplified elevation difference value is selected as the output value, and the contrast at the boundary of the resulting elevation difference image will be more obvious.

When considering the altitude difference between the neighborhood points and the center point, the altitude difference changes of the points closer to the center point can better reflect the overall flatness of the neighborhood. Therefore, the weight values of the points closer to the center point should be increased. The formula with the added distance weight is as follows:

$$g_{x,y} = max\,(\frac{1}{M}\sum_i \gamma_{1i}\cdot\frac{|Z_i - Z_0|}{\sqrt{(X_i - X_0)^2 + (Y_i - Y_0)^2}}$$
$$\frac{1}{M}\sum_i \gamma_{2i}\cdot\frac{|Z_i - Z_0|}{\sqrt{(X_i - X_0)^2 + (Y_i - Y_0)^2}}\Bigg)$$

Where, X_0, Y_0, Z_0 respectively represent the X, Y, and Z-axis values of the LiDAR point projected onto the point (x, y), and (X_i, Y_i, Z_i) represent the X, Y, and Z-axis values of other points in the neighborhood of the center point (x, y). In our refined approach for calculating altitude differences, the inverse of the distance between a certain LiDAR point and the center point is incorporated. This modification places greater emphasis on the contribution of points closer to the center, making their altitude

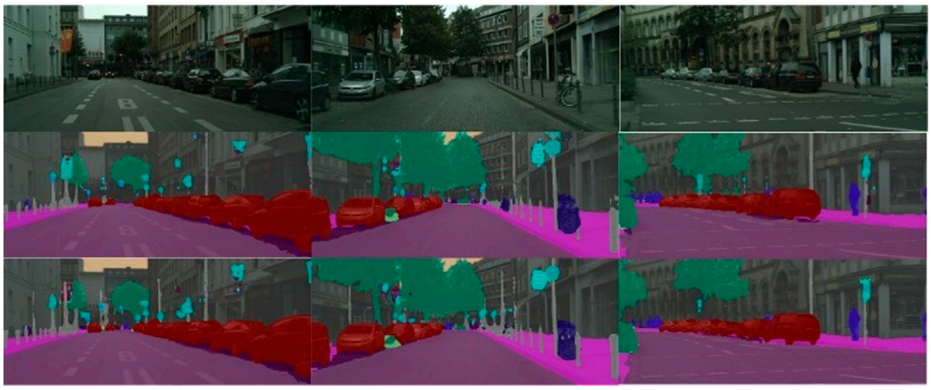


FIGURE 5
Pavement recognition results before and after optimization.

TABLE 1 Perception algorithm accuracy statistics results.

	MaxF (%)	AP (%)	PRE (%)	REC (%)
Image	87.90	90.92	86.66	89.18
Image + WADM	89.39	91.18	88.91	89.87
Image + WADM + FAFM	92.34	92.61	92.65	92.04

differences more pronounced. This technique effectively enhances the distinction between road and non-road areas in the altitude difference image. The impact on road surface points is minimal, preventing any significant intensification in the overall image intensity, while markedly increasing the visibility of non-road surface areas.

For the conversion of point cloud data, we set a 5 × 5 grid centered around (x, y) as the neighborhood range for each point. Consequently, the maximum number of LiDAR points, M, required for computation within this neighborhood is 24 (excluding the center point itself). The algorithm's computational complexity is a function of the generated weighted height map's dimensions (length W and width H), as well as the number of neighborhood points, M. As a result, the computational demand remains low, ensuring the algorithm's real-time performance efficacy. Figure 2 illustrates the outcome of this process: the first row depicts the original RGB image, the second row shows the LiDAR point cloud data, and the third row presents the adaptive weighted altitude difference image. This transformation process converts the initially unordered and sparse point cloud information into a structured, regular two-dimensional image format, where each pixel's grayscale value corresponds to the weighted altitude difference at that location.

3 Feature adaptive fusion network

To integrate the transformed 3D point cloud data with visual image data for better road surface recognition results, we designed a dual-source feature adaptive fusion network, as shown in Figure 3.

The diverse input data sources within our network contribute to a notable disparity between features extracted from the altitude difference map and those derived from visual images. This disparity presents a challenge to the effective fusion of LiDAR and vision

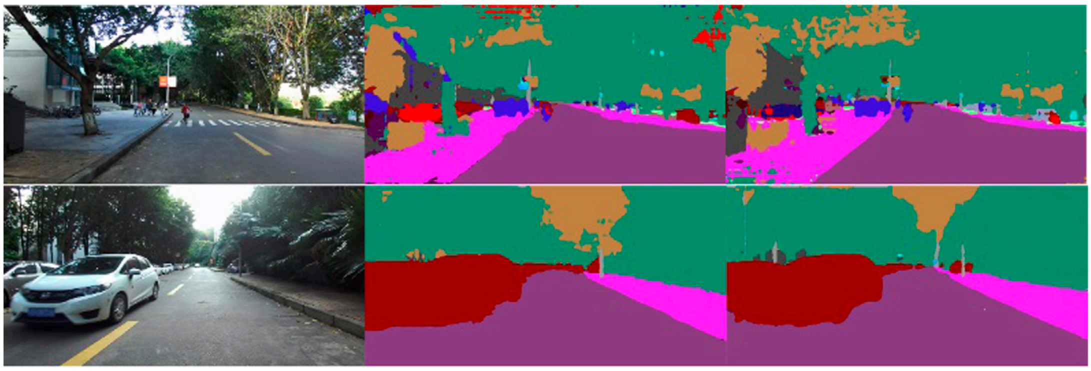


FIGURE 6
Road perception results before and after optimization on real data.

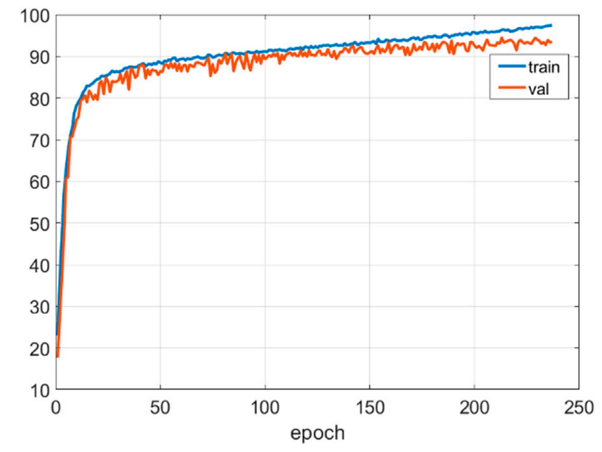


FIGURE 7
Training process diagram. The figure shows the changes in AP during the training process. The model quickly converged to a relatively high level after 50 epochs, and finally completed training after about 240 epochs.

features, hindering seamless integration. To address this challenge, we have devised a methodology for refining features extracted from LiDAR point cloud data. This refinement process enhances the compatibility and synergistic enhancement of LiDAR features with visual features, consequently bolstering road perception performance based on visual inputs.

To materialize this approach, we have developed the Feature Adaptive Fusion Module (FAFM), a novel component comprising two essential elements: the Feature Transformation Network (FTN) and a multi-channel feature weighting cascaded network. The FTN is specifically engineered to adapt LiDAR-derived features to align more cohesively with visual features, facilitating a smoother integration process. Meanwhile, the multi-channel network orchestrates the weighted amalgamation of these refined features. The overarching architecture, illustrated in Figure 4, delineates a sophisticated system that harmoniously leverages the strengths of both LiDAR and visual data for superior road perception capabilities.

3.1 Feature transformation network

The primary objective of the Feature Transformation Network (FTN) is to conduct a linear transformation of LiDAR-derived features, generating new features that exhibit similarity and compatibility with visual image features. This linear transformation is achieved through the following formula:

$$f_{FTN}\left(F_{lidar}\right) = \alpha F_{lidar} + \beta$$

Where, F_{lidar} represents the lidar features, α represents the weight, and β represents the offset. To estimate α and β reasonably and achieve a better fusion of the two features, this paper introduces a feature transformation network to learn and adapt to the lidar features. The following feature transformation network is used to estimate α and β:

$$\alpha = f_{\alpha}\left(F_{lidar}, F_{image}; W_{\alpha}\right)$$
$$\beta = f_{\beta}\left(F_{lidar}, F_{image}; W_{\beta}\right)$$

F_{image} represents the visual image features, f_{α} represents the network function that calculates α, and f_{β} represents the network function that calculates β. W_{α} and W_{β} are the weight parameters of the corresponding networks. The weight values W_{α} and W_{β} are constantly updated during the entire network training process, which makes the estimated weight α and offset β more reasonable.

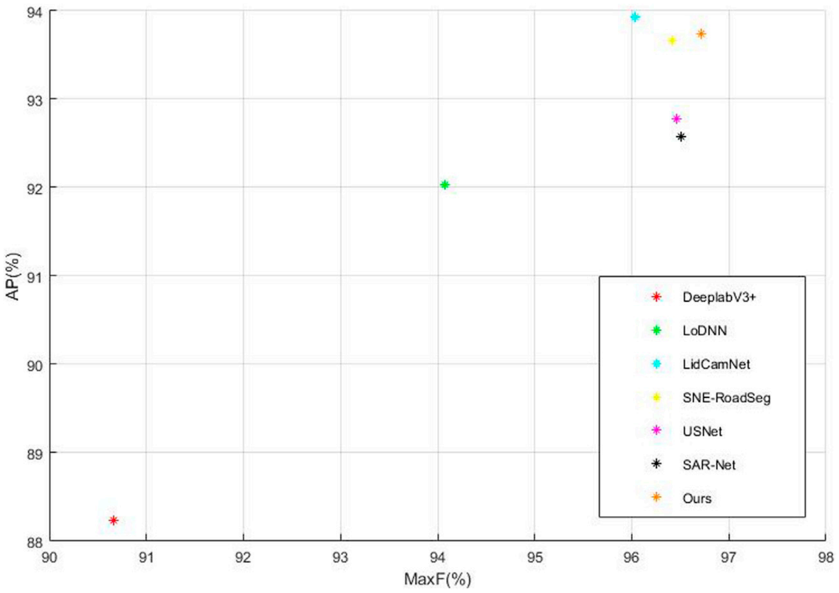


FIGURE 8
Comparison of different algorithms. We used MaxF and AP, the two most significant parameters, as comparison metrics. Our algorithm exhibited a considerable advantage in MaxF and achieved a second-best performance in AP.

TABLE 2 Statistical results of lane extraction accuracy evaluation parameters.

Methods	Input	MaxF (%)	AP (%)	PRE (%)	REC (%)
DeeplabV3+ [12]	Image	90.66	88.23	90.81	90.51
LoDNN [13]	LiDAR	94.07	92.03	92.81	95.37
LidCamNet [14]	Image + LiDAR	96.03	**93.93**	96.23	95.83
SNE-RoadSeg, [15]	Image + LiDAR	96.42	93.67	96.59	96.26
USNet [16]	Image + LiDAR	96.46	92.78	96.32	96.6
SAR-Net [17]	Image + LiDAR	96.51	92.57	**97.36**	96.66
Ours	Image + LiDAR	**96.72**	93.74	96.76	**96.68**

The bolded data represent the best results among the comparison algorithms.

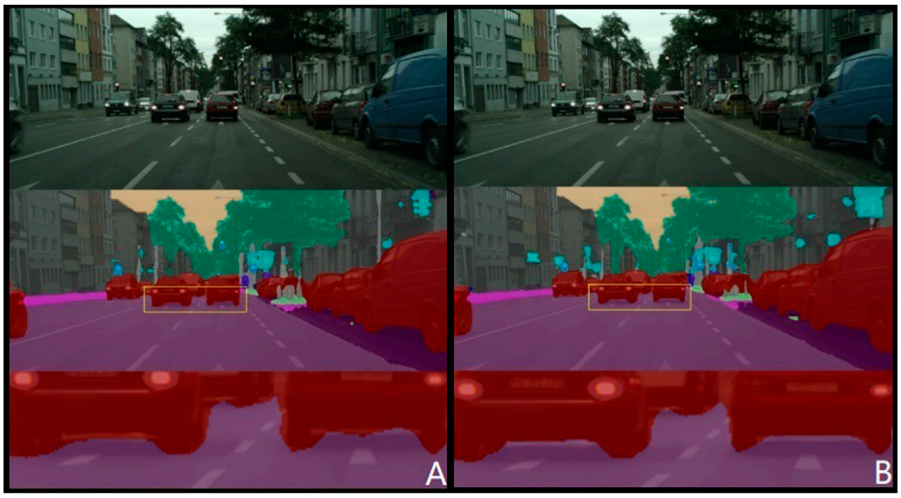


FIGURE 9
Comparison of lane boundary recognition effects. **(A)** DeepLabV3 **(B)** Ours.

The number of output channels for each layer is unified to 256. F_{lidar} and F_{image} are input into the transformation network and their channels are stacked. Two 1 × 1 convolution kernels are used in the transformation network to implement f_α and f_β. The stacked input of F_{lidar} and F_{image} channels is used as input because the 1 × 1 convolution kernel does not change the size of the input feature map. The output is a 256-dimensional weight vector and an offset vector. To avoid introducing too much computational burden, no activation function is added in the transformation network. On the other hand, because the expression ability of the linear model is not sufficient, $(\alpha + 1)$ is selected as the final weight vector to introduce nonlinear factors into the network.

3.2 Multi-channel feature weighted cascade network

The fusion function is achieved by taking the visual image features and the transformed lidar features as inputs, as shown below:

$$f^k_{fuse} = F^k_{image} + \lambda f^k_{FTN}\left(F^k_{lidar}\right)$$

In the context of the road detection system, let k denote the features from the *k*th convolution stage of the Deep Convolutional Neural Network (DCNN), and λ represent a weight parameter. Semantic segmentation heavily relies on information provided by visual image features, with added lidar point cloud features serving as supplementary data. However, experiments have demonstrated that an excessively large proportion of lidar point cloud features can impact the expression of image features, leading to a reduction in semantic segmentation accuracy. Conversely, when the proportion of lidar point cloud features is too small, the effect on algorithmic accuracy optimization is not significant. Optimal balance is achieved when the value of λ is approximately 0.1, resulting in the highest accuracy (subsequent experiments were conducted under the condition of λ = 0.1).

4 Experiments and results

This paper's experimental evaluation comprises two distinct parts: 1) assessing the efficacy of fusing point cloud altitude

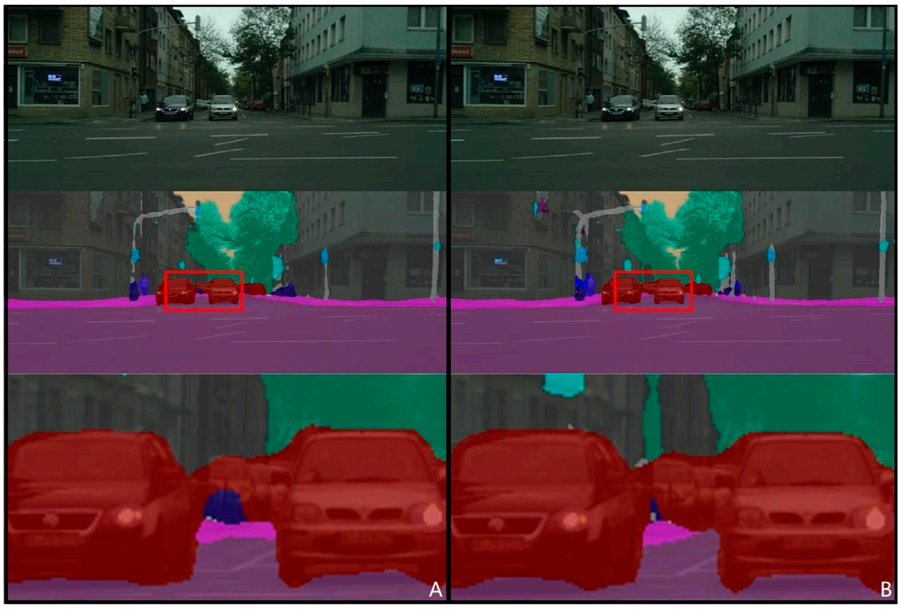


FIGURE 10
Comparison of lane and sidewalk segmentation effects. **(A)** DeepLabV3 **(B)** Ours.

difference data with the feature-adaptive module; 2) benchmarking the recognition accuracy against other leading road detection algorithms.

(1) In the first part, we conducted quantitative assessments of our algorithm's enhancement in road perception accuracy on the public KITTI dataset. We configured three distinct network structures for this purpose: 1) Image: inputs only the visual image, representing the baseline unoptimized network; 2) Image + WADM (Weighted Altitude difference Map): combines the visual image with the adaptive weighted altitude difference map; 3) Image + WADM + FAFM: integrates the visual image and the adaptive weighted altitude difference map, incorporating the feature-adaptive fusion network for a fully optimized algorithm.

As depicted in Figure 5, the results before and after optimization reveal notable differences. The unoptimized road perception algorithm shows marginally weaker semantic segmentation, influenced more significantly by shadows and background luminosity. However, the optimizations, specifically the altitude difference conversion and feature-adaptive fusion, markedly enhance segmentation accuracy. These optimizations address semantic segmentation blurring due to shadows and object occlusion, improving the delineation of segmentation boundaries and the accuracy of distant object perception. Additionally, the integration of LiDAR data bolsters the segmentation effects across various environmental objects.

We further analyzed the performance enhancement of the altitude difference weighted transformation and feature adaptive fusion network. Comparative experiments were conducted under three scenarios, with statistical analyses of various performance metrics tabulated in Table 1. The results affirm that both improvements substantially optimize the algorithm. We used parameters such as MaxF, AP, PRE, and REC to evaluate the algorithm. Their meanings are as follows: MaxF stands for Maximum F1-measure; AP refers to Average Precision as used in PASCAL VOC challenges; PRE indicates Precision; and REC denotes Recall. Notably, the Image + WADM network configuration enhanced the MaxF by 1.49% compared to the baseline, underscoring the significant impact of incorporating LiDAR point cloud information. This addition also positively influenced other parameters, evidencing the improved robustness of the algorithm. The final algorithm model (Image + WADM + FAFM) exhibited the best performance overall, with notable advancements in recall rate and a more balanced performance across all parameters. This underscores the effectiveness and necessity of the feature-adaptive fusion network, confirming its pivotal role in enhancing the algorithm's overall robustness.

In addition, we tested the road perception accuracy of the algorithm before and after full optimization in a real environment. In Figure 6, the first column shows the original visual image (a) in the input network, the second column shows the road perception result under the Image condition (b), and the third column shows the road result under the Image + WADM + FAFM condition (c).

The road perception algorithm designed in this study performs well on both simple and complex structured roads. Compared with the algorithm before optimization, the proposed improvement scheme has improved the accuracy of the algorithm perception and has better robustness under different road conditions. The lane segmentation results are more detailed.

(2) In the lane boundary recognition accuracy experiment, the efficacy of our proposed algorithm was benchmarked against other leading algorithms on the KITTI road dataset. The training process is shown in Figure 7. As detailed in Figure 8; Table 2, our algorithm demonstrates substantial improvements across all accuracy parameters. However, it's noteworthy that the incorporation of two DCNN networks and the fusion network has resulted in a decrease in algorithm speed.

When comparing specific inputs, the LoDNN network, which solely relies on point cloud data, and the DeeplabV3+, which only uses image data, both fall short in overall accuracy compared to algorithms that integrate Image + LiDAR inputs. Among algorithms that employ visual image and LiDAR point cloud data fusion, including LidCamNet, SNE-RoadSeg, USNet, SARNet, and our proposed algorithm, ours shows superior performance in MaxF, PRE, and REC parameters. Although it slightly lags behind LidCamNet in the AP parameter, it maintains a competitive edge.

Based on the subjective and objective evaluation indicators of comprehensive road perception and lane extraction, it can be proved that the algorithm proposed in this paper not only takes into account the effect of road perception, but also has high-precision lane extraction capability.

Our proposed up-sampling network, an enhancement of the Deeplabv3+ network, underwent comparative experiments with the original network. The detailed results, as shown in Figure 9, highlight the algorithm's proficiency. The original image data, road perception results, and lane boundary details are sequentially presented. The proposed algorithm excels at delineating the intersection between lanes and other objects, yielding more precise lane extraction results. This improvement is attributed to the addition of lane edge constraints when converting LiDAR point cloud data into a weighted altitude difference map. This enhancement clarifies lane edge features, heightening their distinctiveness from other objects and facilitating the network's ability to extract the lane area, thereby improving lane recognition accuracy.

In Figure 10, a comparative analysis of segmentation results between two algorithms for lanes and sidewalks underscores our algorithm's superior detection capabilities, even with distant objects. It achieves precise segmentation of lanes and sidewalks, thus significantly enhancing the accuracy of road segmentation at extended distances.

5 Summary

In this study, we meticulously preprocessed the LiDAR point cloud data by removing noise points and optimizing the information within the cloud. This refined 3D point cloud was then projected onto the image plane using specific calibration parameters. A pivotal method based on weighted altitude difference was developed for converting the LiDAR point cloud data. This technique harnessed the height consistency characteristic of flat road areas to extract an altitude difference map from the LiDAR-derived height map. We integrated neighborhood point distance constraints and road boundary point constraints, culminating in the formation of a detailed weighted height map. This innovative approach transforms 3D point cloud data into 2D weighted height map data, adeptly preserving road surface characteristics and accentuating road boundary features. This transformation lays a solid foundation for subsequent fusion with visual imagery. The incorporation of spatial point coordinate information in the point cloud data, coupled with boundary constraints during the conversion process, enabled the explicit representation of road boundary features. This enhancement made the delineation between road and non-road areas more pronounced, greatly benefiting the feature extraction capabilities of subsequent semantic segmentation networks. Additionally, the weighted altitude difference map addresses the susceptibility of visual images to lighting and shadow effects. It remains effective even under challenging conditions of strong light and shadow occlusion, consistently conveying comprehensive road information. The integration of this weighted altitude difference map has significantly bolstered the accuracy of our road perception algorithm, marking a substantial advancement in the field.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cvlibs.net/datasets/kitti/eval_road.php Road/Lane Detection Evaluation 2013.

Author contributions

BW: Conceptualization, Data curation, Formal Analysis, Investigation, Resources, Software, Writing–original draft. ZL: Data curation, Formal Analysis, Resources, Writing–review and editing. XC: Funding acquisition, Supervision, Writing–original draft, Writing–review and editing. CF: Formal Analysis, Funding acquisition, Supervision, Writing–review and editing. MZ: Formal Analysis, Funding acquisition, Writing–review and editing. DS: Writing–review and editing. WW: Formal Analysis, Methodology, Writing–review and editing. SW: Funding acquisition, Project administration, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The National Natural Science Foundation of China (Grant No. 62273063) the Science and Technology Innovation Key R&D Program of Chongqing, China (Grant No. CSTB2022TIAD-STX0003) the China National Foreign Experts Program (G2022165017L).

Conflict of interest

Authors BW and XC were employed by China Merchants Auto-trans Technology Co., Ltd. Authors BW, XC, and CF were employed by China Merchants Testing Vehicle Technology Research Institute Co., Ltd. Author WW was employed by Chongqing Expressway Group Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhang W. Lidar-based road and road-edge detection. In: IEEE Intelligent Vehicles Symposium. IEEE; June, 2010; La Jolla, CA, USA (2010). p. 845–8.

2. Chen T, Dai B, Wang R, Liu D. Gaussian-process-based real-time ground segmentation for autonomous land vehicles. *J Intell Robotic Syst* (2014) 76(3): 563–82. doi:10.1007/s10846-013-9889-4

3. Asvadi A, Premebida C, Peixoto P, Nunes U. 3D Lidar-based static and moving obstacle detection in driving environments: an approach based on voxels and multi-region ground planes. *Robotics Autonomous Syst* (2016) 83(83):299–311. doi:10.1016/j.robot.2016.06.007

4. Wijesoma WS, Kodagoda KRS, Balasuriya AP. Road-boundary detection and tracking using ladar sensing. *IEEE Trans robotics automation* (2004) 20(3):456–64. doi:10.1109/tra.2004.825269

5. Chen L, Yang J, Kong H. Lidar-histogram for fast road and obstacle detection. In: IEEE international conference on robotics and automation (ICRA); May, 2017; Singapore (2017). p. 1343–8.

6. Thrun S, Montemerlo M, Dahlkamp H, Stavens D, Aron A, Diebel J, et al. Stanley: the robot that won the DARPA grand challenge. *J field Robotics* (2006) 23(9):661–92. doi:10.1002/rob.20147

7. Gu S, Zhang Y, Yang J, Kong H. Lidar-based urban road detection by histograms of normalized inverse depths and line scanning. In: 2017 European Conference on Mobile Robots (ECMR); September, 2017; Paris, France (2017). p. 1–6.

8. Fernandes R, Premebida C, Peixoto P, Wolf D, Nunes U. Road detection using high resolution lidar. In: 2014 IEEE Vehicle Power and Propulsion Conference (VPPC); October, 2014; Coimbra, Portugal (2014). p. 1–6.

9. Caltagirone L, Bellone M, Svensson L, Wahde M. LIDAR–camera fusion for road detection using fully convolutional neural networks. *Robotics Autonomous Syst* (2019) 111(111):125–31. doi:10.1016/j.robot.2018.11.002

10. Han X, Lu J, Zhao C, Li H. Fully convolutional neural networks for road detection with multiple cues integration. In: 2018 IEEE International Conference on Robotics and Automation (ICRA); May, 2018; Brisbane, QLD, Australia (2018). p. 4608–13.

11. Liu H, Yao Y, Sun Z, Li X, Jia K, Tang Z. Road segmentation with image-LiDAR data fusion in deep neural network. *Multimedia Tools Appl* (2020) 79(47):35503–18. doi:10.1007/s11042-019-07870-0

12. Chen LC, Papandreou G, Schroff F, Adam H. *Rethinking atrous convolution for semantic image segmentation[J]* (2017). doi:10.48550/arXiv.1706.05587

13. Caltagirone L, Svensson L, Wahde M, Sanfridson M. Lidar-camera Co-training for semi- supervised road detection[J] (2019). doi:10.48550/arXiv.1911.12597

14. Gu S, Yang J, Kong H, A cascaded LiDAR-camera fusion network for road detection. In: 2021 IEEE International Conference on Robotics and Automation (ICRA); May, 2021; Xi'an, China (2021).

15. Fan R, Wang H, Cai P, Liu M, SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection. In: European Conference on Computer Vision; October, 2020; Tel Aviv, Israel (2020). p. 340–56.

16. Chang Y, Xue F, Sheng F, Liang W, Ming A, Fast road segmentation via uncertainty-aware symmetric network. In: IEEE International Conference on Robotics and Automation (ICRA); May, 2022; Philadelphia, PA, USA (2022).

17. Lin H, Liu Z, Cheang C, Xue X. *SAR-net: shape alignment and recovery network for category-level 6D object pose and size estimation[J]* (2021). doi:10.48550/arXiv.2106.14193

18. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille A, Deeplab: semantic image segmentation with deep convolutional nets,atrous convolution, and fully connected CRFS. *IEEE Trans Pattern Anal Machine Intelligence* (2017) 40:834–48. doi:10.1109/tpami.2017.2699184

19. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, et al. Understanding convolution for semantic segmentation. In: Proceedingsof the 2018 IEEE Winter Conference on Applications of Computer Vision(WACV); March, 2018; Lake Tahoe, NV, USA. p. 1451–60.

20. Xiang T, Zhang C, Song Y, Yu J, Cai W. Walk in the cloud: learningcurves for point clouds shape analysis. In: Proceedings of the IEEE/CVF International Conference onComputer Vision (ICCV); October, 2021; Montreal, QC, Canada (2021). p. 915–24.

21. Yu X, Tang L, Rao Y, Huang T, Zhou J, Lu J. Point-bert: pre-training 3dpoint cloud transformers with masked point modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June, 2022; New Orleans, LA, USA (2022).

22. Zhao H, Jiang L, Jia J, Torr PHS, Koltun V. Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); October, 2021; Montreal, BC, Canada (2021). p. 16259–68.

frontiers | Frontiers in Physics

TYPE Review
PUBLISHED 01 July 2024
DOI 10.3389/fphy.2024.1398393

Check for updates

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Wei Wang,
Chengdu University of Information Technology, China
Congyao Zhagn,
Sichuan University, China
Yang Xu,
Chongqing University of Posts and Telecommunications, China
Aleksandar Vakanski,
University of Idaho, United States

*CORRESPONDENCE
Jing Feng,
fjfloating89@126.com

RECEIVED 09 March 2024
ACCEPTED 03 June 2024
PUBLISHED 01 July 2024

CITATION
Song K, Feng J and Chen D (2024), A survey on deep learning in medical ultrasound imaging.
Front. Phys. 12:1398393.
doi: 10.3389/fphy.2024.1398393

A survey on deep learning in medical ultrasound imaging

Ke Song[1], Jing Feng[2]* and Duo Chen[1]

[1]School of Artificial Intelligence, Chongqing University of Education, Chongqing, China, [2]School of Pharmacy (School of Traditional Chinese Medicine), Chongqing Medical and Pharmaceutical College, Chongqing, China

Ultrasound imaging has a history of several decades. With its non-invasive, low-cost advantages, this technology has been widely used in medicine and there have been many significant breakthroughs in ultrasound imaging. Even so, there are still some drawbacks. Therefore, some novel image reconstruction and image analysis algorithms have been proposed to solve these problems. Although these new solutions have some effects, many of them introduce some other side effects, such as high computational complexity in beamforming. At the same time, the usage requirements of medical ultrasound equipment are relatively high, and it is not very user-friendly for inexperienced beginners. As artificial intelligence technology advances, some researchers have initiated efforts to deploy deep learning to address challenges in ultrasound imaging, such as reducing computational complexity in adaptive beamforming and aiding novices in image acquisition. In this survey, we are about to explore the application of deep learning in medical ultrasound imaging, spanning from image reconstruction to clinical diagnosis.

KEYWORDS
medical ultrasound imaging, deep learning, ultrasound beamforming, medical image analysis, clinical diagnosis

1 Introduction

1.1 Brief introduction to medical imaging

Medical imaging relies on various physical phenomena to visualize human body tissues, internally and externally, through non-invasive or invasive techniques. Key modalities such as computed tomography (CT), magnetic resonance imaging (MRI), X-ray radiography, ultrasound, and digital pathology generate essential healthcare data, constituting around 90% of medical information [1]. Consequently, medical imaging plays a vital role in clinical assessment and healthcare interventions. Deep learning, as the cornerstone technology propelling the ongoing artificial intelligence (AI) revolution, exhibits significant potential in medical imaging. It spans from image reconstruction to comprehensive image analysis[2–8]. The integration of deep learning with medical imaging has spurred advancements, with the potential to reshape clinical practices and healthcare delivery. Empirical evidence has proven that deep learning algorithms exhibit performance comparable to that of medical professionals in diagnosing various medical conditions from imaging data [9]. At the same time, many applications of deep learning in clinics have emerged [10–16]. Consequently, there is a discernible trend towards certifying software applications for clinical utilization [17].

1.2 Literature reviews of deep learning in ultrasound beamforming

The development of medical ultrasound has a history of 80–90 years now. Medical ultrasound began as an investigative technology around the end of World War II [18]. With advancements in electronics, this technology improved. Ultrasound is being continually refined for better resolution, more portable devices, and more automated systems that can aid even in remote diagnostics. The most recent advancement in medical ultrasound is the incorporation of AI to help diagnosis.

The application of AI in medical ultrasound is long-standing [19–25]. With the explosion of deep learning, its application in medicine has become even more widespread. The medical ultrasound system mainly includes image reconstruction and image analysis, both of which have seen extensive applications of deep learning [26]. Deep learning has brought a revolutionary change in ultrasound beamforming, significantly enhancing image quality and improving computational efficiency. Ultrasound beamforming is a process of combining signals from multiple ultrasound elements to construct a focused image. Traditional methods rely heavily on user intervention and predefined parameters, which may limit the image quality and accuracy. Deep learning, on the other hand, uses neural network models to learn and generalize from examples. In the context of ultrasound beamforming, deep learning methods can learn to extract relevant features from raw ultrasound data and form a high-quality image without needing explicit instructions or predefined parameters. The process is relatively autonomous and adaptable. In training phase, a deep learning model is trained with a large amount of data (usually raw Radio Frequency (RF) data) which includes both inputs (ultrasound signals) and outputs (desired images). The model learns to identify patterns in the data and how to predict the output from given inputs. Once trained, the model can be used with new input data to predict the corresponding output images. The advantage is that this prediction process is usually faster than traditional beamforming methods as it bypasses the need for complex signal processing. Deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been successfully used in ultrasound beamforming. They have shown promising results in enhancing image resolution, reducing speckle noise, improving contrast, and even performing advanced tasks like tissue characterization and acoustic aberration correction. Around 2017, applications of deep learning in beamforming began to appear in publications [27,28], and the interest in this area has been increasing ever since. In plane wave imaging, if only one plane wave is emitted, a very high frame rate can be achieved, but this will lead to poor image quality. Therefore, to improve image quality, a method called coherent plane wave compounding (CPWC) [29] has been proposed to solve this problem. However, using this method usually requires the emission of plane waves at multiple angles, which leads to a reduction in frame rate. Gasse et al. [27] propose a method using CNNs that allows for the acquisition of high-quality images even with the emission of only three plane waves. Luchies and Byram [28,30] discuss how to use deep neural networks (DNNs) to suppress off-axis scattering. The study is based on operations in the frequency domain through short-time Fourier transform. There are also some studies on bypassing beamforming [31–36]. The principal concept involves utilizing advanced deep learning methodologies to directly reconstruct images or conduct image segmentation from raw RF data. Deep learning has also been used to reduce artifacts in multi-line acquisition (MLA) and multi-line transmission (MLT) [37,38,39]. Luijten et al. [35,40] investigate how deep learning can be applied to the adaptive beamforming process, addressing the computational challenges and aiming to produce better ultrasound images. Wiacek et al. [35,42] explore the use of DNNs to estimate normalized cross-correlation as a function of spatial lag. This estimation is specifically for coherence-based beamforming, such as short-lag spatial coherence (SLSC) beamforming [44]. Using sub-sampled RF data to reconstruct images can increase the frame rate, but the image quality will decrease. Some researchers [31,45–47] propose using deep learning to address this issue. More research is focused on the application of deep learning in plane wave imaging [48–56]. Some studies [57–59] discuss the training schemes. In addition, the ultrasound community also organized a challenge to encourage researchers to engage in deep learning research [60,61].

1.3 Overview of deep learning in clinical application of ultrasound

Deep learning plays a significant role in ultrasound clinical applications as it enhances the efficiency and accuracy of diagnosis, reducing human errors and paving the way for more sophisticated applications. Deep learning models can be trained to automatically detect and segment lesions in ultrasound images. This reduces the workload for radiologists and increases accuracy, as human interpretation can be subjective and variable. They can also be trained to classify diseases based on ultrasound images. Deep learning helps build more detailed 3D and 4D imaging from 2D ultrasound images, providing a more comprehensive picture of the patient’s condition. Deep learning algorithms can be used to predict clinical outcomes or progression of a disease based on ultrasound imaging data. From Ref. [19–25], it can be seen that the application of AI in medical ultrasound analysis predates that of beamforming. Medical ultrasound analysis mainly includes segmentation, classification, registration, and localization [62,63]. The integration of deep learning with ultrasound image analysis has spurred advancements, with the potential to reshape clinical practices. Breast cancer is a disease that seriously threatens people’s health [64]. The application of deep learning in breast ultrasound can effectively assist radiologists or clinicians in diagnosis. Becker et al. [65] are attempting to use a deep learning software (DLS) to classify breast cancer from ultrasound images. Xu et al. [66] focus on segmenting breast ultrasound images into functional tissues using CNNs. This segmentation aids in tumor localization, breast density measurement, and treatment response assessment, crucial for breast cancer diagnosis. Qian et al. [67] discuss a deep-learning system designed to predict Breast Imaging Reporting and Data System (BI-RADS) scores for breast cancer using multimodal breast-ultrasound images. Chen et al. [68] introduce a novel deep learning model for breast cancer diagnosis using contrast-enhanced ultrasound (CEUS) videos. Jabeen et al. [69] present a novel framework for classifying breast cancer from ultrasound images. The method employs deep learning and optimizes feature selection and fusion for enhanced classification accuracy. Raza et al. [70] propose a deep learning framework, DeepBreastCancerNet, designed for the detection and classification of breast cancer from ultrasound images. Deep learning is also widely applied in cardiac ultrasound. Degel et al. [71] discuss a novel approach to segment the left atrium in 3D echocardiography images using CNNs. Leclerc et al. [72] evaluate encoder-decoder deep CNN methods for assessing 2D echocardiographic images. The study introduces the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset, the largest publicly available and fully annotated dataset for echocardiographic assessment, featuring images from 500 patients. Ghorbani et al. [73] investigate the application of deep learning models, particularly CNNs, to interpret echocardiograms. Narang et al. [74] explore the efficacy of a deep learning algorithm in assisting novice operators to obtain diagnostic-quality transthoracic echocardiograms. Ultrasonography is also a primary diagnostic method for thyroid diseases [75]. Thus, some studies on the assistance of deep learning in the diagnosis of thyroid diseases [76–81] emerged. There are also many applications of deep learning in prostate cancer detection [82–84] and prostate segmentation [85–91]. In ultrasound fetal imaging, deep learning also plays an increasingly important role[92–97]. In addition, there is a constant emergence of deep learning research in ultrasound brain imaging [98–103].

1.4 Other review articles on deep learning in ultrasound imaging

There are already some review articles about deep learning in medical ultrasound imaging. van Sloun et al. [104] presents an inclusive examination of the potential and application of deep learning strategies in ultrasound systems, spanning from the front end to more complex applications. In another article [105], they specifically discussed deep learning in beamforming. They introduce the potential role that deep learning can play in beamforming, as well as some of the existing achievements of deep learning in beamforming, and also look forward to new opportunities. Ref. [106] discusses the shortcomings of traditional signal processing methods in ultrasound imaging. The paper suggests a blend of model-based signal processing methods with machine learning approaches, stating that probability theory can seamlessly bridge the gap between conventional strategies and modern machine/deep learning approaches. However, these articles mainly focus on the principles of ultrasound imaging and do not discuss clinical applications. There are also some articles that provide reviews from the perspective of image analysis and clinical practices. Reference [62,63] discuss deep learning in medical ultrasound analysis from multiple perspectives. Afrin et al. [107] discuss the application of deep learning in different ultrasound methods for breast cancer management - from diagnosis to prognosis. Reference [108] presents an in-depth analysis of the application of AI in echocardiography interpretation. Khachnaoui et al. [109] discuss the role of ultrasound imaging in diagnosing thyroid lesions. In this review, our aim is to introduce the application of deep learning in medical ultrasound from the perspective of image reconstruction to clinical applications. The content seems to be quite broad, we aim to provide a comprehensive perspective on the application of deep learning in medical ultrasound and introduce the potential role of deep learning in ultrasound imaging from a system perspective.

2 Overview of medical ultrasound system

A medical ultrasound system consists of various interconnected modules, each of which is further segmented into numerous smaller components. Figure 1 illustrates a simplified block diagram of a medical ultrasound system. The entire signal processing pipeline of an ultrasound system is relatively complex, with even more detailed subdivisions for each module. For those interested in a deeper exploration, please refer to [110]. Here, we are only providing readers with a high-level overview, and a more in-depth introduction to the modules we are interested in will be covered subsequently. The dashed box in Figure 1 represents the analog signal processing module, which is not within the scope of discussion in this survey. We will focus on the discussion of the transmit and receive beamforming and introduce the post-processing as well. The transmit and receive beamforming are actually two distinct parts; however, in Figure 1, we categorize both under beamforming. In subsequent discussions, we will address these two parts separately. The categorization of post-processing here may be overly broad. In fact, after beamforming, there is a series of intermediate processing steps before the final post-processing. However, in this context, we refer to all these processing steps collectively as post-processing.

2.1 Transmit processing

In ultrasound systems, transmit beamforming is a technique that involves controlling the timing of excitation of multiple transducer elements to produce a directional beam or focus the beam within a specific area. By precisely adjusting the phase and amplitude of each element, an ultrasound beam with a specific direction and focal depth can be formed [111].

It can be seen from Figure 2, which illustrates the process of transmit beamforming, that the distance from each element on the transducer to the focal point is different. To ensure the beam ultimately focus at the point, it is necessary to control the emission timing of each element. It is noteworthy that in the typical transmission focusing, only a subset of elements are involved. In plane wave imaging, all elements on the transducer need to transmit, and the direction of the plane wave is controlled by adjusting the transmission timing of each element.

2.2 Receive processing

The receive beamforming is a crucial signal processing technique used to construct a high quality image from the echoes

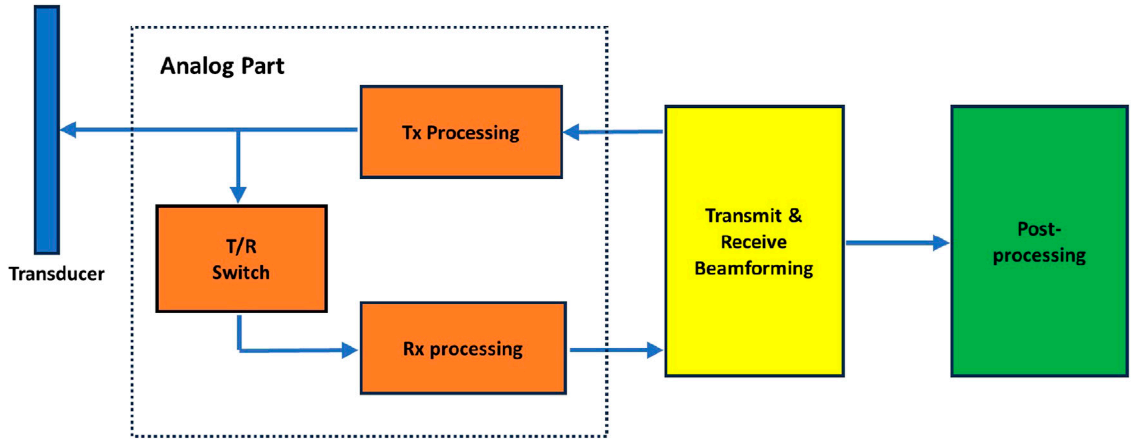


FIGURE 1
Schematic of ultrasound imaging system.

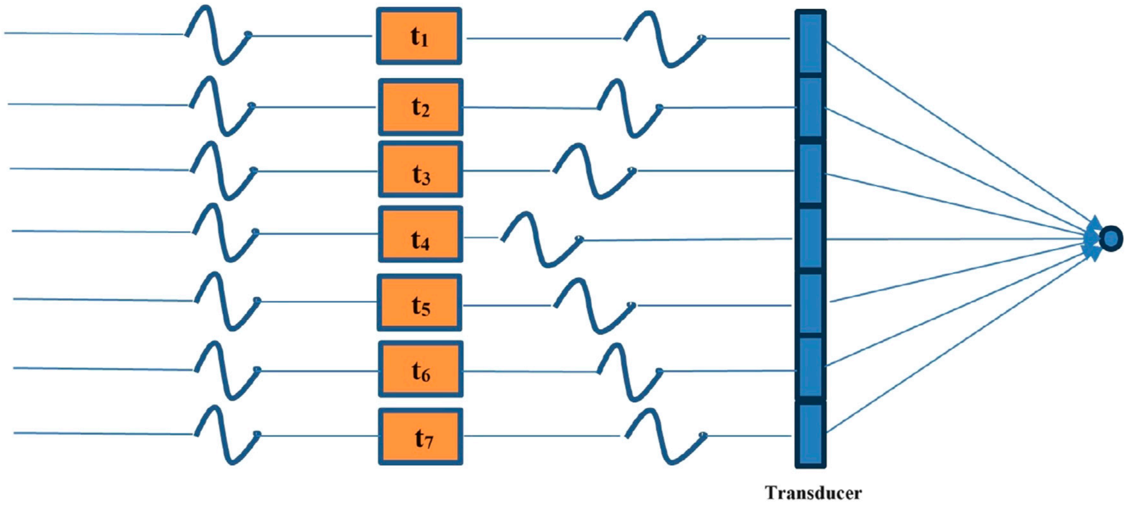


FIGURE 2
Schematic of transmit beamforming. By controlling the emission time of each element on the transducer, the waves emitted by each element can ultimately be focused on one point.

returning from the scanned tissue or organs in ultrasound imaging. When an ultrasound probe emits high-frequency sound waves, they travel through the body, echolocate off structures, and are then reflected back to the receiver. The reflected echoes are captured by multiple transducer elements arranged in an array on the probe. The schematic is illustrated in Figure 3.

Receive beamforming involves combining the signals received by each of these elements in an intelligent way to construct a coherent and high-resolution representation of the scanned region. The most fundamental beamforming technique is the delay-and-sum (DAS) [112] method where the received signals from different transducer elements are delayed relative to each other to account for the different times of flight from the reflecting structure. They are then summed together, enhancing the signal from a specific direction or focal point while attenuating the signals from other directions. While the transmitted beam can be focused at a certain depth, receive beamforming allows dynamic focusing at various depths on receive. The delays are continuously adjusted as the echoes return from different depths, effectively focusing the beam at multiple depths in real-time. The apodization process involves weighting the received signals before they are summed, reducing side lobes and improving the lateral resolution. Advanced beamforming techniques use adaptive methods like Minimum Variance (MV) [113] to improve the image quality further by adapting to the signal environment, hence reducing the impact of off-axis scattering and noise. The result of receive beamforming is a narrow, well-defined beam that can accurately locate and display the internal structures of the body, thus providing detailed images for diagnosis. Advances in digital signal processing and hardware technology have significantly improved beamforming techniques, making them more sophisticated and effective.

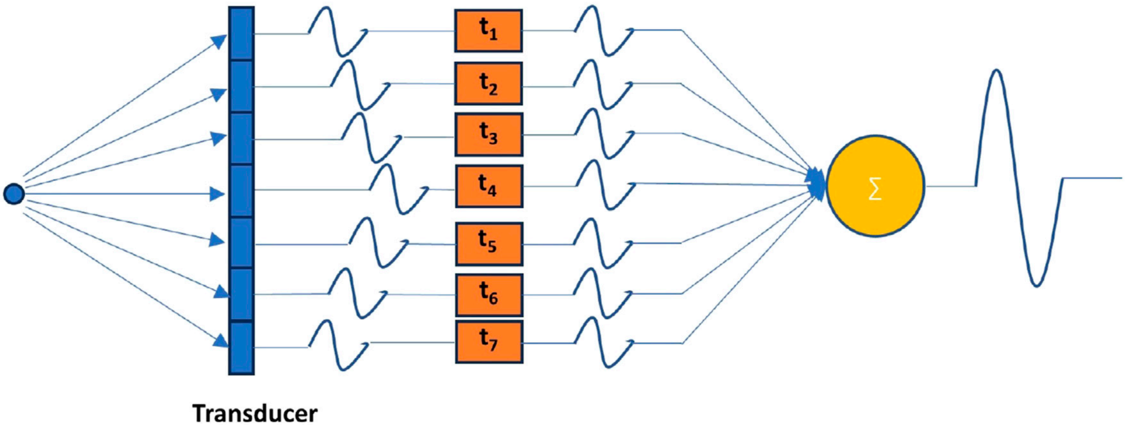


FIGURE 3
Schematic of receive beamforming. In order to align the echo received by each transducer element that is reflected from a certain point, it is necessary to properly delay the signal received by each element.

2.3 Post processing

Ultrasound imaging can be roughly divided into pre-processing and post-processing [114]. Beamforming, as a key part of pre-processing, plays an important role in imaging quality, but post-processing is also an indispensable step. The post-processing is a research field that involves applying several steps after the channel data are mapped to the image domain via beamforming. These steps include further image processing to improve B-mode image quality, such as contrast, resolution, despeckling. It also involves spatiotemporal processing to suppress tissue clutter and to estimate motion. For 2D or 3D ultrasound data, post-processing is crucial for automatic analysis and/or quantitative measurements [115]. For instance, the recovery of quantitative volume parameters is a unique way of making objective, reproducible, and operator-independent diagnoses.

Medical ultrasound image analysis involves the use of diagnostic techniques, primarily those leveraging ultrasound, to create an image of internal body structures like blood vessels, joints, muscles, tendons, and internal organs. These images can then be used to measure certain characteristics such as distances and velocities. Medical ultrasound image analysis has extensive applications in various medical fields, including fetal, cardiac, trans-rectal, and intra-vascular examinations.

Common practices in the analysis of medical ultrasound images often encompass techniques such as segmentation and classification. Segmentation separates different types of organs and structures in the image, especially for regions of interest. Segmentation often uses edge detection, region growing, thresholding techniques, and more advanced techniques such as cascade classifiers, random forests, deep learning, etc. Classification is also a key part of image analysis. It classifies the images into normal images and abnormal images based on the previously extracted features, or further, performs disease classification. Common classification methods include neural networks, K-nearest neighbors (K-NN), decision trees, Support Vector Machines (SVM), etc. In recent years, deep learning-based classification models, such as CNNs, have been widely used, and with their powerful performance and accuracy, are extensively applied in the field of medical image analysis.

3 Deep learning in medical ultrasound imaging

We will discuss the application of deep learning in medical ultrasound imaging from several perspectives. First is the improvement of different beamforming techniques via deep learning, followed by a discussion on clinical application, and then the analysis of the application of deep learning in portable ultrasound devices and training schemes. Finally, we will briefly introduce the CNNs and transformer.

3.1 Image reconstruction

3.1.1 Bypass beamforming

Beamforming plays a crucial role in enhancing image quality. DAS algorithm, as a classic beamforming technique, is widely used in ultrasound imaging systems. Despite its operational simplicity and ease of implementation, this method also presents certain limitations and drawbacks. DAS beamforming generates relatively high side lobes and grating lobes, which are unwanted beam directions that may capture reflected signals from non-target areas, reducing image contrast and resolution. To suppress the side lobes, a common method is to use apodization, which is the application of weighting windows.

Usually, beamforming synthesizes the signals received by an array of elements to form a directional response or beam pattern, but this process can be computationally intensive. Deep learning approaches can potentially learn to perform the beamforming operation more efficiently, leading to faster image reconstruction without compromising quality. Simson et al. [31] address the challenge of reconstructing high-quality ultrasound images from sub-sampled raw data. Traditional beamforming methods, although adept at generating high-resolution images, impose considerable

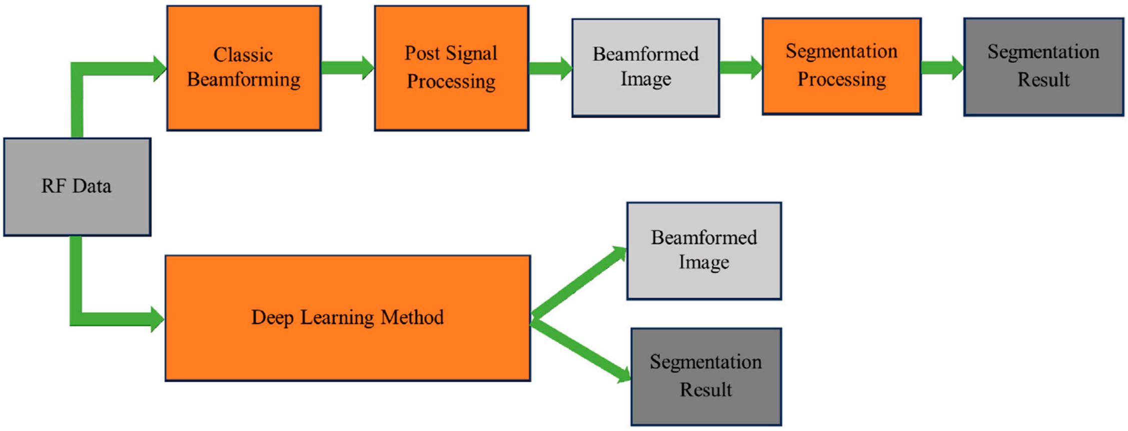


FIGURE 4
A comparison between DAS beamforming (top) and a deep learning method that bypasses the beamforming process [35] (bottom).

computational demands and their efficacy diminishes when dealing with sub-sampled data. To overcome this issue, the authors propose "DeepFormer," an end-to-end, deep learning-based method designed to reconstruct high-quality ultrasound images in real-time, using sub-sampled raw data. Traditional beamforming algorithms often ignore the information between scan lines. Yet, a fully convolutional neural network (FCNN) is capable of capturing this information and utilizing it effectively; thus, enabling cross-scan line interpolation in sub-sampled data. As shown in Eq. 1 [31], the loss function used in DeepFormer is a combination of ℓ_1 loss and Structural Similarity Imaging Metric (SSIM) [116].

$$\mathcal{L}_{DF} = \alpha\mathcal{L}_{MS-SSIM} + (1-\alpha)\mathcal{L}_1 \tag{1}$$

Their results, which were tested on an *in vivo* dataset of some participants, indicate that DeepFormer is a promising approach for enhancing ultrasound image quality while also providing the speed necessary for clinical use. In addition, Nair et al. [32–35], introduced a concept with the objective of achieving high frame rates for automated imaging tasks over an extended field of view using single plane wave transmissions. They address the typical challenge of suboptimal image quality produced by single plane wave insonification and propose the use of DNNs to directly extract information from raw RF data to generate both an image and a segmentation map simultaneously. Unlike traditional beamforming, which generally only reconstructs images, they have utilized deep learning to achieve both image reconstruction and segmentation at the same time. They employed FCNN, the entire network includes an encoder and two decoders, one for image reconstruction and the other for image segmentation. As shown in Eq. 2 [35], the loss function of the entire network also adopts a combination of ℓ_1 loss and Dice similarity coefficient (DSC) loss.

$$L_T(\theta) = \ell_1(\theta) + DSC(\theta) \tag{2}$$

The DSC loss is used to measure the overlap between the predicted and true segmentation masks during the training of their DNN. Specifically, the DSC loss is utilized to quantify the similarity between the predicted DNN segmentation and the true segmentation. The DSC is calculated as a function of the overlap between these two segmentations, with a value of one indicating perfect overlap and 0 indicating no overlap. The DSC loss complements the mean absolute error loss by focusing on the segmentation performance of the network. While the mean absolute error provides a pixel-wise comparison between the predicted and reference images, the DSC loss offers a more holistic measure of the segmentation quality, especially important in medical imaging where the precise localization of structures is vital. This dual-loss approach enables the network to learn both the image reconstruction and segmentation tasks effectively, ensuring that the network parameters are optimized to generate accurate segmentations alongside the reconstructed images. The comparison between classic beamforming and this method is shown in Figure 4.

3.1.2 Adaptive beamforming

Traditional beamforming typically uses a fixed, predetermined set of weights applied to the received signals from each transducer element. These weights are usually uniform (DAS) or they use simple apodization (windowing) techniques. The resolution is generally limited by the fixed nature of the weights. The main lobe width does not adapt to different signal scenarios, which can lead to a less focused image. Traditional approaches may exhibit relatively higher side lobes, inducing higher levels of interference and clutter within the image. However, these methods are simpler to implement and faster in terms of computation, which makes them suitable for many real-time imaging applications.

The MV beamforming uses an adaptive approach to determine the weights applied to the signals. It calculates the weights that minimize the variance of the noise and interference, essentially optimizing the signal-to-noise ratio. The adaption of weights allows to generate a much narrower main lobe in the beam pattern, which translates to higher spatial resolution and better ability to distinguish between closely spaced scatterers. As a result of the narrower main lobe and suppressed side lobes, MV can provide significantly improved image resolution and contrast. It allows for clearer delineation of structures within the body, especially beneficial when visualizing small or closely spaced scatterers. The MV algorithm uses the data from the transducer elements to estimate the covariance matrix of the received signals. As shown in Eq. 3

[113], the weights are derived to minimize the output variance while maintaining the gain in the direction of the signal of interest.

$$\begin{aligned} \hat{\mathbf{w}} &= \min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_x \mathbf{w} \\ s.t. \quad & \mathbf{w}^H \mathbf{a} = 1 \end{aligned} \tag{3}$$

where $\mathbf{R}_x$ is the covariance matrix of the received signals and $\mathbf{a}$ is a steering vector of ones. This optimization process, typically solved as a constrained minimization problem, is more complex than applying fixed weights as in traditional beamforming methods. This process involves calculating the correlations between the signals received at each pair of array elements. As the number of elements increases, the size of this matrix grows quadratically, thus increasing the computational burden. To compute the weights that will minimize the variance of the noise and interference, the MV algorithm requires the inversion of the covariance matrix. The inversion of a matrix is considered a process that requires significant computational resources, especially as the size of the matrix grows with the number of transducer elements. The algorithm must dynamically adapt and recalculate the weights for each focal point in real-time as the transducer moves and steers its beam. This continuous adaptation requires the algorithm to perform the above computations for each new set of received signals, which is computationally demanding.

Luijten et al. [35,40] examine the applicability of deep learning to augment the adaptive beamforming process, addressing the computational challenges and aiming to produce better ultrasound images. They develop a neural network architecture, termed Adaptive Beamforming by deep LEarning (ABLE), which can adaptively calculate apodization weights for image reconstruction from received RF data. This method aims to improve ultrasound image quality by efficiently mimicking adaptive beamforming methods without the high computational burden. The ABLE network consists of fully connected layers and employs an encoder-decoder structure to create a compact representation of the data, aiding in noise suppression and signal representation. The training of ABLE is performed using a specialized loss function designed to promote similarity between the target and the produced images while also encouraging unity gain in the apodization weights. The study demonstrates ABLE's effectiveness on two different ultrasound imaging modalities: plane wave imaging with a linear array and synthetic aperture imaging with a circular array. Moreover, ABLE's computational efficiency, as assessed by the number of required floating-point operations, is significantly lower than that of Eigen-Based Minimum Variance (EBMV) beamforming, highlighting its potential for real-time imaging applications. In the training strategy, the network employs a total loss function composed of an image loss and an apodization-weight penalty. The image loss is designed to promote similarity between the target image and the one produced by ABLE, while the weight penalty encourages the network to learn weights that facilitate a distortionless response in the beamforming process. This penalty is inspired by MV beamforming principles, which aim to minimize output power while ensuring a distortionless response in the desired direction. By incorporating this constraint, the network is guided to learn apodization weights that not only aim to reconstruct high-quality ultrasound images but also adhere to a fundamental beamforming criterion, ensuring the network's predictions align with the physical beamforming process. The comparison between MV and ABLE is shown in Figure 5.

3.1.3 Spatial coherence-based beamforming

Spatial coherence-based beamforming is a sophisticated method employed in ultrasound imaging that focuses on analyzing the spatial coherence of received echo signals to form diagnostic images. It improves image clarity by emphasizing echoes that show consistent phase or time delays across neighboring transducer elements, which indicates they are coming from a real reflector-like tissue structure, rather than random noise or scattering. By harnessing this spatial coherence, the beamformer can more effectively differentiate between signal and noise, leading to images with better resolution and contrast.

Typically, the DAS algorithm only utilizes one attribute, the signal strength, while spatial coherence reflects the similarity of signals [117]. Therefore, this is another property that can be used to enhance image quality. There are many studies based on spatial coherence, such as coherence factor (CF) [118], generalized coherence factor (GCF) [119], and phase coherence factor (PCF) [120]. Lediju et al. [44] have proposed a spatial coherence-based method named short-lag spatial coherence (SLSC). This method leverages the coherence of echoes that occur at short lags. The objective of this method is to overcome the limitations of traditional ultrasound imaging, caused by factors such as acoustic clutter, speckle noise, and phase aberration. SLSC images demonstrate improved visualization when compared to matched B-mode images by addressing these issues. By applying the SLSC imaging, the researchers aim to enhance ultrasound image quality and diagnostic accuracy, benefiting the field of medical imaging. The spatial coherence is calculated by Eq. 4 [44],

$$\hat{R}(m) = \frac{1}{N-m} \sum_{i=1}^{N-m} \frac{\sum_{s=s_1}^{s_2} x_i(s) x_{i+m}(s)}{\sqrt{\sum_{s=s_1}^{s_2} x_i^2(s) \sum_{s=s_1}^{s_2} x_{i+m}^2(s)}} \tag{4}$$

where x_i is the aligned signal received by the ith element, s_i represents the sample index along the axial direction. In addition, N denotes the receive aperture, and m indicates the lag. From this equation, it can be seen that its computational complexity is relatively high. Wiacek et al. [35,42] have proposed a deep learning approach named CohereNet to estimate the normalized cross correlation as a function of lag. This network can be used to replace the SLSC beamforming. They delve into the potential of FCNNs as "universal approximators" that could learn any function. In CohereNet, a 7 × 64 input is adopted, which means the axial kernel chooses seven samples in the axial direction, while the aperture size is 64. The output is the spatial correlation at different lag distances. The network structure consists of an input layer, three fully connected layers using rectified linear unit (ReLU) as the activation function, followed by a fully connected layer using hyperbolic tangent (tanh) as the activation function, and an average pool output layer. In essence, CohereNet aims to utilize the capability of DNNs to enhance the beamforming process, thereby improving image quality and computational efficiency. As described in [43], the CohereNet is faster than SLSC, and this network also has high generality. The Figure 6 illustrates the comparison between DAS and CohereNet.

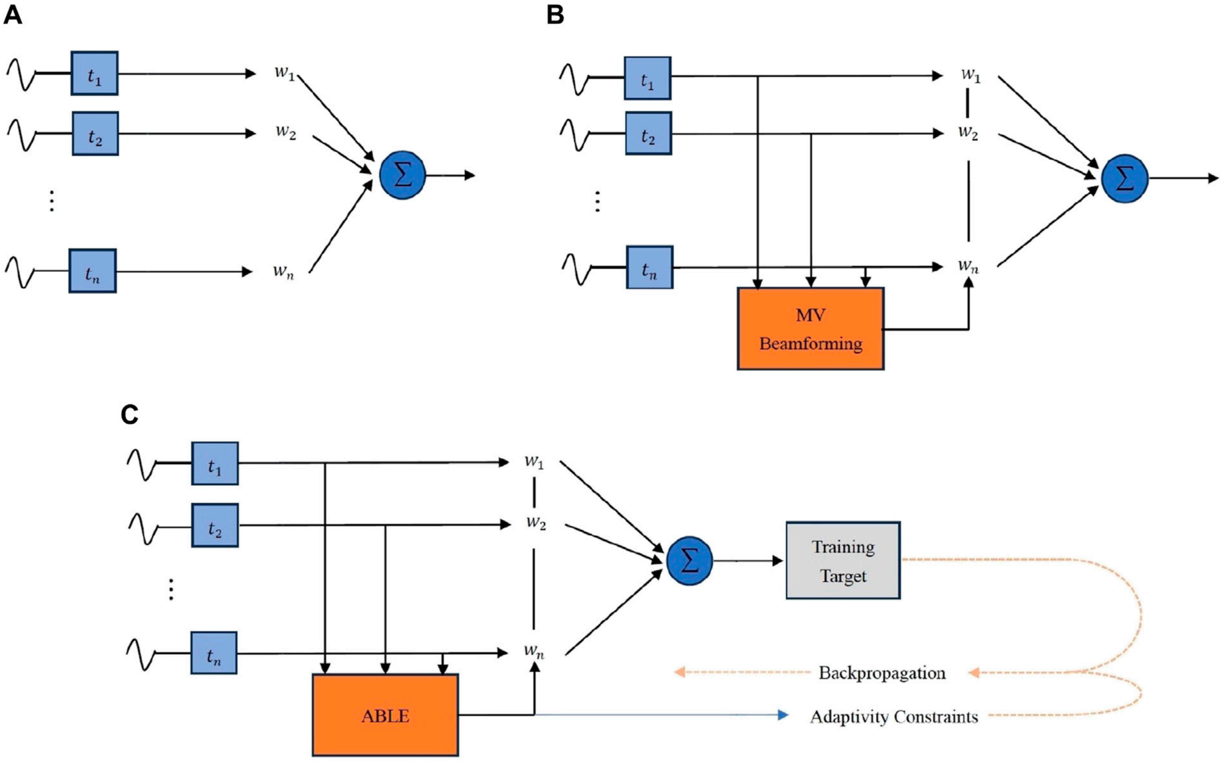


FIGURE 5
A comparison among **(A)**DAS, **(B)**MV and **(C)**ABLE [41]. The weights in DAS are typically pre-set fixed values, while the weights in MV and ABLE are adaptive. ABLE can be seen as an alternative form of MV. They both adaptively estimate weights through the received signals. The calculation of weights in MV requires a large amount of computation, while ABLE reduces the computational complexity.

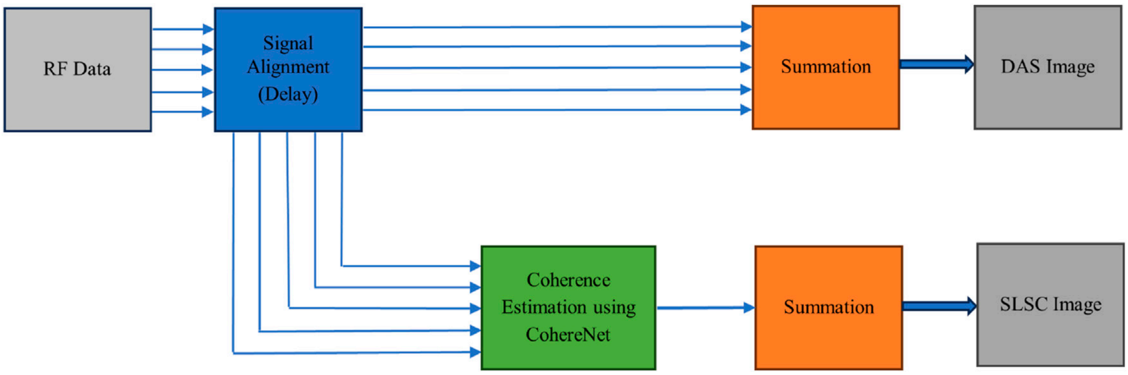


FIGURE 6
A comparison between DAS beamforming (top) and CohereNet [43] (bottom). The DAS algorithm obtains the final result by aligning the received signals and then weighting and summing them up. On the other hand, SLSC achieves the final result through calculating the spatial coherence. CohereNet reduces the computational complexity of SLSC.

3.2 Deep learning in clinical applications

Ultrasound imaging, due to its non-invasive characteristic and real-time imaging capabilities, has seen extensive use across various medical domains. This section delves into the clinical applications of deep learning, including breast imaging, cardiology, prostate imaging, fetal, thyroid, and brain.

3.2.1 Breast imaging

Breast ultrasound imaging is commonly utilized to detect potential breast diseases [121]. Although it falls short in identifying microcalcifications compared to X-ray mammography, it is instrumental in distinguishing benign masses like cysts and fibroadenomas from malignant ones. With the development of AI, especially the advent of deep learning

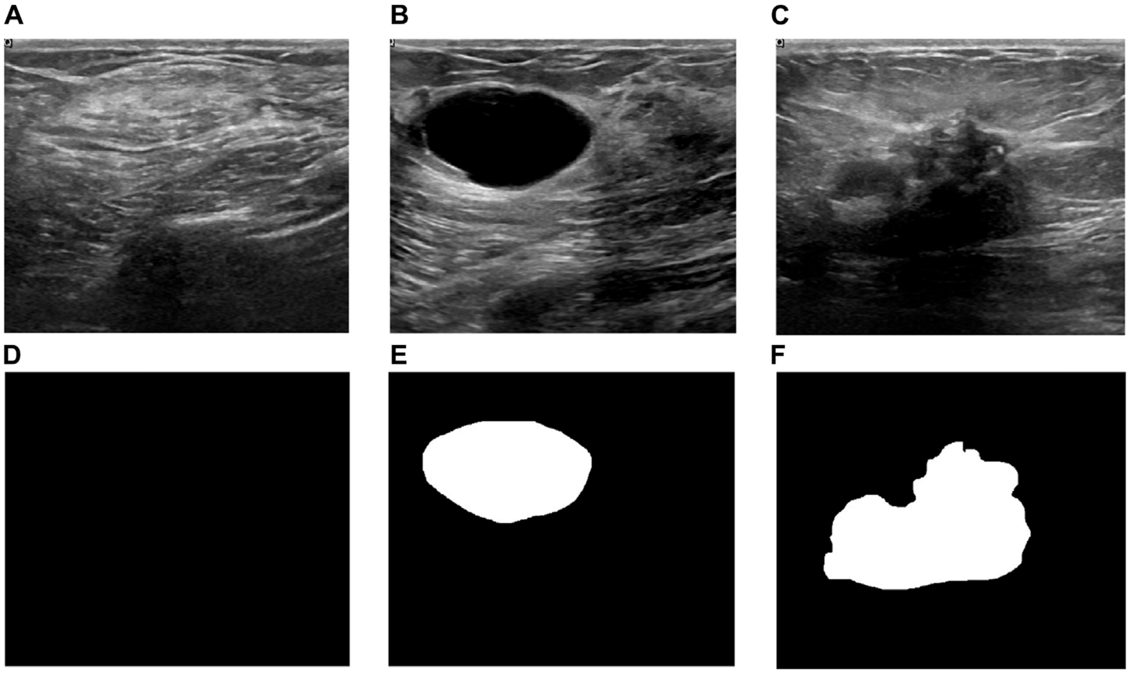


FIGURE 7
Three examples from BUSI dataset[122]. **(A)** A normal image and **(D)** its corresponding mask, **(B)** a benign image and **(E)** its corresponding mask, **(C)**a malignant image and **(F)** its corresponding mask.

technologies, it has also promoted the evolution of ultrasound breast imaging. Some open-source datasets, such as Breast Ultrasound Images Dataset (BUSI) [122], have also promoted the widespread application of deep learning. As depicted in Figure 7, the images within the BUSI dataset are classified into three distinct categories: normal, benign, and malignant.

In 2018, Becker et al.[65] reported using generic deep learning software (DLS) for the classification of breast cancer in ultrasound images. The study aimed to evaluate the effectiveness of a DLS in classifying breast cancer using ultrasound images and compare its performance against human readers with varying levels of breast imaging experience. They used Receiver Operating Characteristic (ROC) to assess the accuracy of diagnostic results. The DLS achieved diagnostic accuracy comparable to radiologists and performed better than a medical student with no prior experience. Although they did not discuss the technical details of deep learning, the study demonstrated that deep learning software could achieve high diagnostic accuracy in classifying breast cancer using ultrasound images, even with a limited number of training cases. The fast evaluation speed of the software supports the feasibility of real-time image analysis during ultrasound examinations. This indirectly illustrates the potential of deep learning in improving diagnostic processes. Xu et al. [66] develop a CNN based method for the automatic segmentation of breast ultrasound images into four major tissues: skin, fibroglandular tissue, mass, and fatty tissue, to aid in tumor localization, breast density measurement, and assessment of treatment response. They designed two CNN architectures CNN-I and CNN-II. CNN-I is an 8-layer CNN for pixel-centric patch classification. CNN-II is a smaller CNN to combine the outputs of three CNN-Is, each trained on orthogonal image planes, to provide comprehensive evaluation. CNNs were trained using the Adam optimization algorithm, and dropout methods were applied to prevent overfitting. The proposed method achieved high quantitative metrics for segmentation. Accuracy, Precision, Recall, and F1-measure all exceeded 80%. Jaccard similarity index (JSI) for mass segmentation reached 85.1%, outperforming previous methods. The proposed method provided better segmentation visualization and quantitative evaluation compared to previous studies. The automated segmentation method can offer objective references for radiologists, aiding in breast cancer diagnosis and breast density assessments. Qian et al. [67] have proposed a deep learning system to assess the breast cancer risk. The system was trained on a large dataset from two hospitals, encompassing 10,815 ultrasound images of 721 biopsy-confirmed lesions, and then prospectively tested on an additional 912 images of 152 lesions. The deep-learning system, when applied to bimodal (B-mode and color Doppler images) and multimodal (including elastography) images, achieved high accuracy in predicting BI-RADS scores. The system's predictions align with radiologists' assessments, demonstrating its potential utility in clinical settings. It could facilitate the adoption of ultrasound in breast cancer screening, particularly beneficial for women with dense breasts where mammography is less effective. This research underscores the potential of deep learning in enhancing breast ultrasound's diagnostic power, offering a tool that aligns with current BI-RADS standards and supports radiologists in decision-making processes. Chen et al. [68] introduce a deep learning model for breast cancer diagnosis. They leverage the domain knowledge of radiologists, particularly their diagnostic patterns when viewing CEUS videos, to enhance the model's diagnostic accuracy. The model integrates a 3D CNN with a domain-knowledge-guided temporal attention module (DKG-TAM) and a domain-knowledge-guided channel attention module (DKG-CAM). These modules are designed to mimic the attention patterns of radiologists, focusing on specific time slots in contrast-enhanced ultrasound (CEUS) videos and incorporating relevant features from both CEUS and traditional ultrasound images. The study utilizes a Breast-CEUS dataset comprising 221 cases, which includes CEUS videos and corresponding images, making it one of the largest datasets of its kind. Reference [69] addresses the challenge of breast cancer, the second leading cause of death among women worldwide. It highlights the importance of early detection through automated systems due to the time-consuming nature of manual diagnosis. The study introduces a new framework leveraging deep learning and feature fusion for classifying breast cancer using ultrasound images. The proposed framework comprises five main steps: data augmentation, model selection, feature extraction, feature optimization, feature fusion and classification. Operations like horizontal flip, vertical flip, and 90-degree rotation were applied to enhance the original dataset's size and diversity. The pre-trained DarkNet-53 model was modified and trained using transfer learning techniques. Features were extracted from the global average pooling layer of the modified model. Two improved optimization algorithms, reformed differential evolution (RDE) and reformed gray wolf (RGW), were used to select the best features. The optimized features were fused using a probability-based approach and classified using machine learning algorithms. The study concludes that the proposed framework significantly improves the accuracy and efficiency of breast cancer classification from ultrasound images. It highlights the potential of the method to provide reliable support for radiologists, enhancing early detection and treatment planning. Rzaz et al. [70] present DeepBreastCancerNet, a new deep learning model designed for the detection and classification of breast cancer using ultrasound images. This model addresses the challenges of manual breast cancer detection, which is often time-consuming and prone to inaccuracies. The proposed DeepBreastCancerNet framework includes 24 layers, consisting of six convolutional layers, nine inception modules, and one fully connected layer. It employs both clipped ReLU and leaky ReLU activation functions, batch normalization, and cross-channel normalization to enhance model performance. Images were augmented through random translations and rotations to enhance the dataset's size and diversity, thereby reducing overfitting. The architecture starts with a convolutional layer followed by max pooling, batch normalization, and leaky ReLU activation. Inception modules are used for extracting multi-scale features. The model ends with a global average pooling layer and a fully connected layer for classification. The proposed model achieved a classification accuracy of 99.35%, outperforming several state-of-the-art deep learning models. On a binary classification dataset, the model achieved an accuracy of 99.63%. The DeepBreastCancerNet model outperformed other pre-trained models like AlexNet, ResNet, and GoogLeNet in terms of accuracy, precision, recall, and F1-score. Ablation studies confirmed the importance of using both leaky ReLU and clipped ReLU activation functions and global average pooling for optimal performance.

3.2.2 Cardiology

In cardiology, echocardiography, particularly through ultrasound imaging of the heart, represents a pivotal area in medical ultrasound research, with abundant literature focusing on automated methods for segmenting and tracking the heart's left ventricle - a crucial component evaluated in heart disease diagnosis. Echocardiography is a test that uses high-frequency sound waves to make pictures of your heart. It can show the size, shape, movement, pumping strength, valves, blood flow and other features of the heart. The quality of echocardiographic images can be influenced by multiple factors such as patient's body habitus, lung disease, or surgical dressings, which can make interpretation difficult. Interpreting the results of an echocardiography exam requires significant expertise and experience. Sometimes not all views of the heart can be visualized adequately, which may limit the amount of information obtained from the test. Deep learning techniques can solve these problems to a certain extent.

Ref.[71] addressed the challenge of segmenting the left atrium (LA) in 3D ultrasound images using CNNs. The proposed method aims to automate this process, which is traditionally time-consuming and dependent on the observer. The introduction of shape priors and adversarial learning into the CNN framework enhances the accuracy and adaptability of the segmentation across different ultrasound devices. The framework integrates three existing methods: 3D Fully Convolutional Segmentation Network (V-Net), Anatomic Constraint via Autoencoder Network and Domain Adaptation with Adversarial Networks. The V-Net processes 3D image volumes and creates segmentation masks. Shape priors are incorporated through an autoencoder network trained on ground truth segmentation masks. This ensures that the segmentation masks adhere to anatomically plausible shapes. Domain adaptation is achieved by training a classifier to identify the data source, aiming to make the feature maps domain invariant. The combined approach of using shape priors and adversarial learning in CNNs significantly improves the segmentation of the left atrium in 3D ultrasound images. This method not only boosts accuracy but also ensures the generalizability of the model across different devices, making it a promising tool for clinical use.

2D echocardiographic image analysis is crucial in clinical settings for diagnosing cardiac morphology and function. Manual and semi-automatic annotations are still common due to the lack of accuracy and reproducibility of fully automatic methods. Challenges in segmentation arise from poor contrast, brightness inhomogeneities, speckle patterns, and anatomical variability. Leclerc et al.[72] evaluate the performance of state-of-the-art encoder-decoder deep CNNs for segmenting cardiac structures in 2D echocardiographic images and estimating clinical indices using the CAMUS dataset. CAMUS is the largest publicly available and fully annotated dataset for echocardiographic assessment, containing data from 500 patients. The CAMUS dataset enables comprehensive evaluation of deep learning methods for echocardiographic image analysis. Encoder-decoder networks, especially U-Net, demonstrate strong potential for accurate and reproducible cardiac segmentation, paving the way for fully automatic analysis in clinical practice. The study confirms that encoder-decoder networks, particularly U-Net, provide highly accurate segmentation results for 2D echocardiographic images. However, achieving inter-observer variability remains challenging, and more sophisticated architectures did not significantly outperform simpler U-Net designs. The findings suggest that further improvements in deep learning methods and larger annotated datasets are essential for advancing fully automatic cardiac image analysis.

Ghorbani et al. [73] have developed a deep learning model named EchoNet to interpret the echocardiograms. This model could identify local cardiac structures, estimate cardiac function, and predict systemic phenotypes like age, sex, weight, and height with significant accuracy. EchoNet is able to accurately predict various clinical parameters, such as ejection fraction and volumes, crucial for diagnosing and managing heart conditions. It also demonstrated the potential to predict systemic phenotypes that are not directly observable in echocardiogram images. By automating echocardiogram interpretation, such AI models could streamline clinical workflows, provide preliminary interpretations in regions lacking specialized cardiologists, and offer insights into phenotypes challenging for human evaluation. The research emphasized the potential of deep learning to enhance echocardiogram analysis, offering a step toward more automated, accurate, and comprehensive cardiovascular imaging diagnostics. Due to the lack of experience among novices, Narang et al. [74] proposed the use of deep learning techniques to assist them. The deep learning algorithm provides real-time guidance to novices, enabling them to capture essential cardiac views without prior experience in ultrasonography. The study involved eight nurses without prior echocardiography experience who used the AI guidance to perform echocardiographic scans on 240 patients. These scans were then compared with those obtained by experienced sonographers. The primary outcome was the ability of the AI-assisted novices to acquire echocardiographic images of sufficient quality to assess left and right ventricular size and function, as well as the presence of pericardial effusion. Results indicated that the novice-operated, AI-assisted echocardiograms were of diagnostic quality in a high percentage of cases, closely aligning with the quality of scans performed by experienced sonographers. The study suggests that AI-guided echocardiogram acquisition can potentially expand the accessibility of echocardiographic diagnostics to settings where expert sonographers are unavailable, thereby enhancing patient care in diverse clinical environments.

3.2.3 Thyroid

The thyroid gland, located in the neck and comprising two interconnected lobes, plays a critical role in hormone secretion, impacting protein synthesis, metabolic rate, and calcium homeostasis. These hormones are particularly influential in children's growth and development. Despite its small size, the thyroid is susceptible to various disorders, such as hyperthyroidism, hypothyroidism, and nodule formation. Diagnosing these conditions involves a range of techniques, including blood tests for hormone levels, ultrasound imaging for gland volume and nodule detection, and fine-needle aspiration (FNA) for definitive tissue analysis. FNA, the most invasive of these methods, is being increasingly circumvented by leveraging ultrasound imaging with advanced deep learning and computer-aided diagnosis (CAD) systems to enhance diagnostic accuracy and nodule characterization.

Wang et al. [79] introduce a deep learning method for diagnosing thyroid nodules using multiple ultrasound images from an examination. The study proposes an architecture that includes three networks, addressing the challenge of using multiple views from an ultrasound examination for a comprehensive diagnosis. The research involves a dataset with 7803 images from 1046 examinations, employing various ultrasound equipment. The dataset is annotated at the examination level, categorizing examinations into malignant and benign based on ultrasound reports and pathological records. The method integrates features from multiple images using an attention-based feature aggregation network, aiming to reflect the diagnostic process of sonographers who consider multiple image views. The model demonstrated high diagnostic performance on the dataset, showcasing the potential of deep learning in enhancing the accuracy and objectivity of thyroid nodule diagnosis in ultrasound imaging. The attention-based network assigns weights to different images within an examination, focusing on those with significant features, which aligns with clinical practices where sonographers prioritize certain image views. Peng et al. [77] have developed a deep learning AI model called ThyNet. This model was designed to differentiate between malignant tumors and benign thyroid nodules, aiming to enhance radiologists' diagnostic performance and reduce unnecessary FNAs. ThyNet was developed using 18,049 images from 8,339 patients across two hospitals and tested on 4,305 images from 2,775 patients across seven hospitals. The model's performance was initially compared with 12 radiologists, and then a ThyNet-assisted diagnostic strategy was developed and tested in real-world clinical settings. The AI model, ThyNet, demonstrated superior diagnostic performance compared to individual radiologists, with an area under the receiver operating characteristic curve (AUROC) of 0.922, significantly higher than the radiologists' AUROC of 0.839. When radiologists were assisted by ThyNet, their diagnostic performance improved significantly. The pooled AUROC increased from 0.837 to 0.875 with ThyNet assistance for image reviews and from 0.862 to 0.873 in a clinical setting involving image and video reviews. The ThyNet-assisted strategy significantly decreased the percentage of unnecessary FNAs from 61.9% to 35.2%, while also reducing the rate of missed malignancies from 18.9% to 17.0%.

3.2.4 Prostate

Prostate cancer ranks as the most frequently diagnosed malignancy among adult and elderly men, with early detection and intervention being crucial for reducing mortality rates. Transrectal ultrasound (TRUS) imaging, in conjunction with prostate-specific antigen (PSA) testing and digital rectal examination (DRE), plays a pivotal role in the diagnosis of prostate cancer. The delineation of prostate volumes and boundaries is critical for the accurate diagnosis, treatment, and follow-up of this cancer [123]. Typically, the delineation process involves outlining prostate boundaries on transverse parallel 2-D slices along its length, leading to the development of various (semi-) automatic methods for detecting these boundaries. In the diagnosis of prostate diseases, deep learning techniques provide some additional insights.

Azizi et al. [83] present a deep learning approach using Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for prostate cancer detection through Temporal Enhanced Ultrasound (TeUS). The study aimed to leverage the temporal information inherent in TeUS to distinguish between malignant and benign tissue in the prostate. The authors utilized RNNs to model the temporal variations in ultrasound backscatter signals, demonstrating that LSTM networks outperformed other models in identifying cancerous tissues. The study analyzed data from 255 prostate biopsy cores from 157 patients. LSTM networks achieved an area under the curve (AUC) of 0.96, with sensitivity, specificity, and accuracy rates of 0.76, 0.98, and 0.93, respectively, highlighting the potential of RNNs in medical imaging analysis. The study also introduced algorithms for analyzing LSTM networks to understand the temporal features relevant for prostate cancer detection. This analysis revealed that significant discriminative features could be captured within the first half of the TeUS sequence, suggesting a potential reduction in data acquisition time for clinical applications. The research suggests that deep learning models, particularly LSTM-based RNNs, can significantly enhance prostate cancer detection using ultrasound imaging, offering a promising tool for improving diagnostic accuracy and potentially guiding biopsy procedures. Karimi et al. [89] introduces a method for the automatic segmentation of the prostate clinical target volume (CTV) in TRUS images, which is crucial for brachytherapy treatment planning. The method employs CNNs, specifically an ensemble of CNNs, to improve segmentation accuracy, particularly for challenging images with weak landmarks or strong artifacts. The method uses adaptive sampling to focus the training process on difficult-to-segment images and an ensemble of CNNs to estimate segmentation uncertainty, improving robustness and accuracy. For segmentations with high uncertainty, a statistical shape model (SSM) is used to refine the segmentation, utilizing prior knowledge about the expected shape of the prostate. The method achieved a Dice score of 93.9% ± 3.5% and a Hausdorff distance of 2.7 ± 2.3 mm, outperforming several other methods and demonstrating its effectiveness in reducing the likelihood of large segmentation errors. This study highlights the potential of deep learning and ensemble methods to enhance the accuracy and reliability of medical image segmentation, particularly in applications like prostate cancer treatment where precision is crucial.

3.2.5 Fetal

Ultrasonography is a pivotal technology in prenatal diagnosis, renowned for its safety for both the mother and fetus. This research area encompasses numerous subfields, often employing segmentation and classification techniques akin to those used in adult diagnostics but adapted for the smaller scale of fetal organs. This miniaturization introduces diagnostic challenges due to less pronounced signs of abnormalities. Furthermore, ultrasound imaging must penetrate maternal tissue and the placenta to reach the fetus, potentially introducing noise, exacerbated by the movements of both mother and fetus, emphasizing the need for enhanced automated diagnostic methods. Especially in underdeveloped areas with a shortage of medical personnel, such automatic diagnostic methods can provide tremendous help.

Van den Heuvel et al. [94] present a study where a system is developed to estimate the fetal head circumference (HC) from ultrasound data obtained using an obstetric sweep protocol (OSP). This protocol can be taught within a day to any healthcare worker without prior knowledge of ultrasound. The study aims to make ultrasound imaging more accessible in developing countries by eliminating the need for a trained sonographer to acquire and interpret images. The system uses two FCNNs. The first network detects frames containing the fetal head from the OSP data, and the second network measures the HC from these frames. The HC measurements are then used to estimate gestational age (GA) using the curve of Hadlock. The study, conducted on data from 183 pregnant women in Ethiopia, found that the system could automatically estimate GA with a reasonable level of accuracy, indicating its potential application in maternal care in resource-constrained settings. Pu et al. [96] developed an automatic fetal ultrasound standard plane recognition (FUSPR) model. This model is designed to operate in an Industrial Internet of Things (IIoT) environment and leverages deep learning to identify standard planes in fetal ultrasound imagery. The research introduces a distributed platform for processing ultrasound data using IIoT and high-performance computing (HPC) technology. The FUSPR model integrates a CNN and an RNN to learn spatial and temporal features of ultrasound video streams, aiming to improve the accuracy and robustness of fetal plane recognition. The system's goal is to aid in gestational age assessment and fetal weight estimation by accurately identifying and analyzing key anatomical structures in ultrasound video frames. The study demonstrates that the FUSPR model significantly outperforms baseline models in recognizing four standard fetal planes from over 1000 ultrasound videos. The use of deep learning within the IIoT framework presents a promising approach to enhancing the efficiency and reliability of fetal ultrasound analysis, particularly in resource-constrained environments. A study by Xu et al.[124] introduced a novel segmentation framework incorporating vector self-attention layers (VSAL) and context aggregation loss (CAL) to address the challenges of fetal ultrasound image segmentation. The VSAL module allows for simultaneous spatial and channel attention, capturing both global and local contextual information. The CAL component further enhances the model's ability to differentiate between similar-looking structures by considering both inter-class and intra-class dependencies. On the multi-target Fetal Apical Four-chamber dataset and one-target Fetal Head dataset, the proposed framework outperformed several state-of-the-art CNN-based, U-net [125], methods in terms of pixel accuracy (PA), dice coefficient (DCS), Hausdorff distance (HD) metrics, demonstrating its potential for improving fetal ultrasound image segmentation accuracy. The study showcases the effectiveness of self-attention techniques in enhancing the accuracy and reliability of fetal ultrasound image segmentation, offering a promising tool for improving prenatal diagnostics and care.

3.2.6 Brain

The brain, a pivotal organ in the nervous system, epitomizes complexity within the human body, orchestrating the functions of voluntary organs and muscles. Despite its critical role, the full extent of its operations remains partially elusive, prompting ongoing research to decipher its mechanisms. Notably, the brain undergoes a phenomenon termed "brain shift," a potential deformation during surgical procedures that could impact surgical outcomes. Ultrasound technology, particularly when integrated with magnetic resonance (MR) imaging data, serves as a crucial aid in neurosurgical contexts. This integration is instrumental in addressing the challenges posed by brain shift and enhancing intraoperative navigation and decision-making.

Milletari et al.[98] discuss a deep learning approach using CNNs combined with a Hough voting strategy for segmenting deep brain regions in MRI and ultrasound images. The study showcases the use of this method for fully automatic localization and segmentation of anatomical regions of interest, utilizing the features produced by CNNs for robust, multi-region, and modality-flexible segmentation. The method is particularly designed to adapt to different imaging modalities, showing effectiveness in MRI and transcranial ultrasound volumes. It demonstrates the potential of CNNs in medical image analysis, particularly in the challenging context of brain imaging, where accurate segmentation of anatomical structures is critical. The study systematically explores the performance of various CNN architectures across different scenarios, offering insights into the effective application of deep learning techniques in medical imaging. Reference [99] presents a method for segmenting brain tumors during surgery using 3D intraoperative ultrasound (iUS) images. The technique employs a tumor model derived from preoperative magnetic resonance (MR) data for local MR-iUS registration, aiming to enhance the visualization of brain tumor contours in iUS. This multi-step process defines a region of interest based on the patient-specific tumor model, extracts hyperechogenic structures from this region in both modalities, and performs registration using gradient values and rigid and affine transformations to align the tumor model with the 3D-iUS data. The method's effectiveness was assessed on a dataset of 33 patients, showing promising results in terms of computational time and accuracy, indicating its potential utility in supporting neurosurgeons during brain tumor resections. Di Ianni and Airan [102] introduce a deep learning-based image reconstruction method for functional ultrasound (fUS) imaging of the brain. The method significantly reduces the amount of data required for imaging while maintaining image quality, using a CNN to reconstruct power Doppler images from sparsely sampled ultrasound data. The approach enables high-quality fUS imaging of brain activity in rodents, with potential applications in various settings where dedicated ultrasound hardware is not available, thereby broadening the accessibility and utility of fUS imaging technology.

3.3 Deep learning in portable ultrasound system

Due to its portability and low cost, handheld ultrasound devices have great application prospects in areas such as emergencies, point-of-care, sports fields, and outdoors. At the same time, it is also suitable for assisting doctors in diagnosing diseases in remote and medically undeveloped areas. Portable ultrasound diagnostic devices appeared in the 1980s. Initially, they were mainly used to scan the bladder to measure the volume [126–129]. Compared to the common invasive method of catheterization through a urinary catheter, the bladder scanner does not cause any harm to the patient. Until now, the development of bladder scanners has been a direction in the advancement of portable ultrasound devices [130–132]. However, besides this field, portable ultrasound devices have many other applications, such as Color Doppler imaging [133], blood flow imaging [134], echocardiography [135], skin imaging [136] and so on. During the outbreak of COVID-19, portable ultrasound devices also played a positive role in assisting diagnosis [137–139].

Indeed, the compact size of portable ultrasound devices does present significant challenges for both hardware design and the development of imaging algorithms [140,141]. Despite its portable advantages, these challenges need to be meticulously addressed to ensure the efficient performance and accuracy of the device. With the advancement of semiconductor technology, technologies such as Field Programmable Gate Arrays (FPGAs) [142] and Application Specific Integrated Circuits (ASICs) [143] have been successively applied to portable ultrasound devices to overcome some of the challenges in hardware design. From the perspective of algorithm design, beamforming technology based on compressed sensing [144] has extensive research in portable ultrasound [145–148].

These methods have promoted the development of portable ultrasound devices, and with the advancement of artificial intelligence technology, the corresponding technologies have also brought new development directions for portable ultrasound devices. Zhou et al. [149] proposed to apply Generative Adversarial Network (GAN) to enhance the image quality of handheld ultrasound devices. They introduce a novel approach using a two-stage GAN to enhance image quality. The proposed two-stage GAN framework incorporates a U-Net network and a GAN to reconstruct high-quality ultrasound images from low-quality ones. The method focuses on reconstructing tissue structure details and speckles of the ultrasound images, essential for accurate diagnostics. The paper presents a comprehensive loss function combining texture, structure, and perceptual features to guide the GAN training effectively. The simulated, phantom and clinical data are used to demonstrate the method's efficacy, showing significant improvements in image quality compared to original low-quality images and other algorithms. In addition, Soleimani et al. [103] developed a lightweight and portable ultrasound computed tomography (USCT) system for noninvasive imaging of the human head with high resolution. The study aims to compare the effectiveness of a deep neural network combining CNN and long short-term memory (LSTM) layers against traditional deterministic methods in creating tomographic images of the human head. The research shows that the proposed neural network is more effective in dealing with noisy and synthetic data compared to deterministic methods, which often require additional filtering to improve image quality. The findings suggest that the CNN + LSTM model is more versatile and generalizable, making it a superior choice for medical ultrasound tomography applications. The study contributes to the advancement of USCT by demonstrating the potential of deep learning approaches in improving the accuracy and reliability of noninvasive brain imaging techniques.

3.4 Training scheme

Vienneau et al. [57] discuss the training methods for DNNs in the context of ultrasound imaging. They address the issues with traditional ℓ_p norm loss functions when training DNNs, where lower loss values do not necessarily translate to improved image quality. Ref. [57] presents an effort to better align the optimization objective with the relevant image quality metrics. The authors suggest that their novel training scheme can potentially increase the maximum achievable image quality for ultrasound beamforming using DNNs.

Luchies and Byram [58] investigate practical considerations of training DNN beamformers for ultrasound imaging. They discuss the use of combinations of multiple point target responses for training DNNs, as opposed to single point target responses. It also examines the impact of various hyperparameter settings on the quality of ultrasound images in simulated scans. The study demonstrates that DNN beamforming exhibits robustness when confronted with electronic noise, and it points out that mean squared error (MSE) validation loss is not a reliable predictor for image quality. Goudarzi and Rivaz [52] used real photographic images as the ground-truth echogenicity map in their simulations to provide the network with a diverse range of textures, contrasts, and object geometries during the training phase. This approach not only enhances the variety in the training dataset, which is crucial for preventing overfitting but also aligns the simulation settings more closely with the real experimental imaging settings of *in vivo* test data, thus minimizing unwanted domain shifts between training and test datasets.

3.5 Transformer/attention mechanism and CNN

CNNs have been the backbone of medical image analysis for years. It can be seen from our previous review that a larger number of architectures are based on the CNNs.They excel in extracting local features through convolutional layers, pooling, and activation functions. Networks such as U-Net[125] and its variants [150–153] have been particularly successful in medical image segmentation tasks due to their encoder-decoder architectures, which capture detailed spatial hierarchies. However, CNNs face limitations in modeling global context and long-range dependencies. This shortfall can lead to suboptimal performance in tasks where the relationship between distant regions in the image is crucial. In ultrasound imaging, this limitation manifests in difficulties handling speckle noise and artifacts, which require broader contextual understanding to be effectively mitigated. On the other hand, the advent of Transformer models and their self-attention mechanisms [154] has introduced new opportunities for enhancing ultrasound image analysis. The integration of U-net with transformer has also become a new direction for current research [155–159]. This section delves into the application of Transformers and attention mechanisms in medical imaging focusing on ultrasound, comparing their performance with traditional CNNs.

Transformers, originally designed for natural language processing, utilize a self-attention mechanism that allows the model to weigh the importance of different input elements dynamically [154]. This capability is particularly beneficial for medical image analysis [124,156,160–167], where different regions of an image may hold varying levels of significance for accurate diagnosis. The self-attention mechanism operates by creating attention scores between all pairs of input elements, which in the context of images, correspond to pixels or features. These scores determine how much attention each element should receive from the others. This global consideration enables the Transformer to capture long-range dependencies and contextual information that CNNs might miss due to their localized receptive fields [154]. In the realm of medical imaging, the Transformer models have been adapted to handle the unique challenges posed by this modality. For instance, TransUNet [156] architecture integrates CNNs and Transformers into a unified framework, where CNNs are employed to extract initial feature maps from medical images, and Transformers encode these features into tokenized patches to capture global context. This hybrid approach enables the model to retain detailed spatial information while benefiting from the global attention provided by Transformers; the GPA-TUNet[162] model integrates Group Parallel Axial Attention (GPA) with Transformers to enhance both local and global feature extraction. This hybrid approach leverages the strengths of Transformers in capturing long-range dependencies and the efficiency of GPA in highlighting local information. Another segmentation method specific to ultrasound images is the integration of a Vector Self-Attention Layer (VSAL) [124], which performs long-range spatial and channel-wise reasoning simultaneously. VSAL is designed to maintain translational equivariance and accommodate multi-scale inputs, which are critical for handling the variability in ultrasound images. This layer can be seamlessly integrated into existing CNN architectures, enhancing their performance by adding the benefits of self-attention. Studies [124] have shown that Transformer-based models significantly improve the accuracy of ultrasound image segmentation tasks. For example, in the segmentation of fetal ultrasound images, models incorporating VSAL and context aggregation loss (CAL) demonstrated superior performance compared to traditional CNNs.

The adaptive multimodal attention mechanism [160] is another advanced approach used in deep learning models to improve the generation of descriptive and coherent medical image reports. Yang et al. propose a novel framework for generating high-quality medical reports from ultrasound images using an adaptive multimodal attention network (AMAnet). This framework addresses the challenges of tedious and time-consuming manual report writing by leveraging deep learning techniques to automate the process. The core innovation of AMAnet lies in its adaptive multimodal attention mechanism, which integrates three key components: spatial attention, semantic attention, and a sentinel gate. The spatial attention mechanism focuses on the relevant regions of the ultrasound images, ensuring that the model captures essential visual details. Meanwhile, the semantic attention mechanism predicts crucial local properties, such as boundary conditions and tumor morphology, by using a multi-label classification network. These predicted properties are then used as semantic features to enhance the report generation process. The sentinel gate is a pivotal element in the AMAnet framework, designed to dynamically control the attention level on visual features and language model memories. This gate allows the model to decide whether to focus on current visual features or rely on the learned knowledge stored in the Long Short-Term Memory (LSTM) when generating the next word in the report. This adaptive mechanism is particularly beneficial in handling fixed phrases commonly found in medical reports, ensuring that the model can generate coherent and contextually appropriate text. The incorporation of semantic features and the adaptive attention mechanism contribute to the model's superior performance, highlighting its potential for practical clinical applications. In practical terms, consider a scenario where the model is generating a report for an ultrasound image showing a tumor. The spatial attention mechanism might focus on the region where the tumor is located. The semantic attention mechanism will consider properties such as "irregular morphology" and "unclear boundary" predicted by the multi-label classification network. The sentinel gate will dynamically balance between these features and the language model's internal memory to generate a sentence like "The ultrasound image shows an irregularly shaped tumor with unclear boundaries." This adaptive attention mechanism ensures that the model generates accurate and contextually appropriate reports, enhancing its utility in clinical settings, which CNNs alone might struggle to achieve.

Chi et al. [168] propose a unified framework that combines the 2D and 3D Transformer-UNets into a single end-to-end network. This novel method enhances the segmentation of thyroid glands in ultrasound sequences, addressing several key limitations of existing deep learning models. The proposed Hybrid Transformer UNet (H-TUNet) integrates both intra-frame and inter-frame features through a combination of 2D and 3D Transformer UNets, significantly improving segmentation accuracy and efficiency. The framework is designed to exploit both the detailed intra-frame features and the broader inter-frame contextual information, resulting in a more accurate and robust segmentation of the thyroid gland in ultrasound images. The proposed method outperforms state-of-the-art CNN-based models, such as 3D UNet, in terms of segmentation accuracy, demonstrating the effectiveness of hybrid Transformer-2D-3D models in ultrasound image analysis. Wang et al.[169] presents a groundbreaking method for enhancing the safety and efficiency of robot-assisted prostate biopsy through advanced force sensing techniques. This method addresses the limitations of existing VFS techniques, particularly in accurately sensing the interaction force between surgical tools and prostate tissue. The core innovation of TransVFS is the spatio-temporal local–global transformer architecture. This model captures both local image details and global dependencies simultaneously, which is crucial for accurately estimating prostate deformations and the resulting forces during biopsy. The architecture includes efficient local–global attention modules that reduce the computational burden associated with processing 4D spatio-temporal data. This makes the method suitable for real-time force-sensing applications in clinical settings. The proposed method was extensively validated through experiments on prostate phantoms and beagle dogs. The results demonstrated that TransVFS outperforms state-of-the-art VFS methods and other spatio-temporal transformer models in terms of force estimation accuracy. Specifically, TransVFS provided significantly lower mean absolute errors in force estimation compared to the most competitive model, ResNet3dGRU. The paper highlights the practical benefits of TransVFS in improving the safety and efficacy of robot-assisted prostate biopsies. By providing accurate real-time force feedback, TransVFS can help reduce the risk of tissue damage and improve the precision of biopsy procedures, thereby enhancing patient outcomes. Ahmadi et al.[170] integrate a spatio-temporal architecture that combines anatomical features and the motion of the aortic valve to accurately classify AS severity. The Temporal Deformable Attention (TDA) mechanism is specifically designed to capture small local motions and spatial changes across frames, which are critical for assessing AS severity. The model incorporates a temporal coherent loss function to enforce sensitivity to small motions in spatially similar frames without explicit aortic valve localization labels. This loss helps the model maintain consistency in frame-level embeddings, enhancing its ability to detect subtle changes in the aortic valve's movement. An innovative attention layer is introduced to aggregate disease severity likelihoods over a sequence of echocardiographic frames, focusing on the most clinically informative frames. This temporal localization mechanism enables the model to identify and prioritize frames that are critical for accurate AS diagnosis. The model was tested on both private and public datasets, demonstrating state-of-the-art accuracy in AS detection and severity classification. On the private dataset, the model achieved 95.2% accuracy in AS detection and 78.1% in severity classification. On the public TMED-2 dataset, the model achieved 91.5% accuracy in AS detection and 83.8% in severity classification. By reducing the reliance on Doppler measurements and enabling automated AS severity assessment from two-dimensional echocardiographic data, the proposed framework facilitates broader access to AS screening. This is particularly valuable in clinical settings with limited access to expert cardiologists and specialized Doppler imaging equipment.

Transformers address the limitations of CNNs by incorporating self-attention mechanisms that consider the entire input sequence (or image) simultaneously. This allows for a more comprehensive understanding of the image, capturing both local and global features effectively. Transformers can capture long-range dependencies and relationships across the entire image, which is essential for accurately interpreting ultrasound images that may contain complex structures and subtle differences. By dynamically adjusting the attention weights, Transformers can focus on the most relevant parts of the image, enhancing feature extraction and reducing the impact of irrelevant or noisy regions. Recent methods further develop the advantages via Integration with CNNs: Hybrid models, such as GPA-TUNet[162], combine the strengths of CNNs and Transformers, using CNN layers for initial feature extraction and Transformers for global context modeling. This integration leads to superior performance in segmentation tasks, particularly for images with large axial spans. Adaptive Attention Mechanisms: Models like AMAnet[160] incorporate adaptive attention mechanisms that dynamically control the focus on visual features and language model memories. This enables the model to generate coherent and contextually appropriate reports, enhancing its utility in clinical settings. Efficient Spatio-Temporal Processing: Methods like TransVFS[169,170] introduce factorized spatio-temporal processing strategies that significantly reduce computational complexity, making them suitable for real-time force-sensing applications in clinical settings. These advanced techniques demonstrate the potential of Transformers and attention mechanisms in enhancing the accuracy and reliability of medical ultrasound image analysis, offering a promising solution for improving diagnostic outcomes and patient care.

The integration of Transformer models and attention mechanisms into ultrasound image analysis represents a significant advancement over traditional CNN-based approaches. The ability of Transformers to capture long-range dependencies and model global context enhances the accuracy and reliability of medical image segmentation tasks. As research continues, these models [160,162,169,170] are likely to play an increasingly vital role in improving diagnostic accuracy and patient outcomes in medical imaging.

4 Discussion

With the rapid development of deep learning, its range of applications has also expanded into more fields. In this paper, we summarize the applications of deep learning in medical ultrasound imaging, focusing on its promoting effect on beamforming algorithms and clinical applications. We compared the classic beamforming algorithm and its corresponding deep learning alternatives. For both adaptive beamforming and SLSC beamforming algorithms, the use of deep learning can reduce computational complexity and enhance efficiency. Deep learning can enhance beamforming algorithms in medical ultrasound imaging in several ways. Data-Driven Optimization: Deep learning models can be trained on large datasets of ultrasound images to learn optimal beamforming parameters for different imaging conditions. This can result in better image quality compared to traditional beamforming techniques that use preset parameters. Feature Extraction: Neural networks, especially CNNs, are highly efficient at automatically extracting relevant features from ultrasound data. These features can then be employed to improve the spatial and contrast resolution of the images. Reducing Artifacts: Deep learning can help identify and reduce artifacts in ultrasound images, such as speckle noise, which can interfere with the clarity of the images and the diagnosis. Speeding Up Processing Time: Deep learning can significantly reduce the computational time required for beamforming, making real-time imaging more feasible and efficient. Advanced Reconstruction Techniques: Through the use of deep learning, more advanced beamforming algorithms, such as synthetic aperture and plane wave imaging, can be optimized for better resolution and frame rates. In summary, deep learning can play a crucial role in the advancement of beamforming algorithms by enhancing image quality, reducing noise, and improving the overall efficiency of medical ultrasound imaging.

In the section on “clinical applications”, we reviewed the application of deep learning in some clinical scenarios. The specific applications of deep learning in medical image analysis include the following aspects. Image registration and orientation: Deep learning can align the spatial orientation and adjust the pixel intensity of multiple images from different sources, times, directions, or modalities to increase the effective sample size and reduce non-biological differences. Tissue segmentation: Deep learning technology can achieve precise segmentation of target structures in medical images, which helps to improve the speed and accuracy of medical image analysis. Disease prediction and diagnosis: Deep learning can assist doctors in diagnosing various diseases, including tumors, inflammations, injuries, etc. For example, it has been successfully used in the diagnosis of many diseases such as lung cancer and breast cancer. Medical image feature learning: Intelligent calculations of medical imaging based on deep learning can automatically learn excellent feature expressions from large sample data.

Deep learning, as an advanced machine learning technique, has significant potential in improving the performance of beamforming algorithms in medical ultrasound. Deep learning may have a positive impact on beamforming algorithms in medical ultrasound in the future. Deep learning can improve the quality of ultrasound images by denoising, enhancing edges and contrast, and reconstructing details more finely. Accelerating the beamforming computational process through deep learning models could significantly reduce the time required to acquire high-quality ultrasound images. Deep learning models can optimize beamforming algorithms based on different patient characteristics and scanning conditions to achieve more personalized imaging. Deep learning models can increase the dynamic range of images, making it possible to display both high and low signal areas in the same image, and enhance resolution. It can also identify and reduce artifacts in ultrasound imaging, such as sidelobe contamination and Doppler artifacts. Beamforming algorithms integrated with deep learning can assist in real-time detection of lesions and measurement of biomarkers, providing more diagnostic information. Deep learning can be used for rapid reconstruction of three-dimensional and four-dimensional data, providing clinicians with a more comprehensive view. Deep learning can help perform more accurate tissue quantitative analysis, such as the measurement of tissue stiffness, which is particularly important for certain diagnoses. By continuously learning from clinical data, deep learning models can improve their performance over time, enhancing the accuracy and reliability of beamforming technology. Deep learning can be used to automatically determine the optimal beamforming parameters, simplifying clinical operations and reducing the workload of physicians.

The integration of AI in portable ultrasound devices with remote servers is also helpful. AI algorithms can analyze ultrasound images in real-time, helping to identify patterns, anomalies, or specific conditions. This can assist healthcare professionals in making more accurate and faster diagnoses. AI can enable remote monitoring of patients, analyzing ultrasound data transmitted to the remote server and alerting healthcare professionals to any concerning changes or findings that require immediate attention. AI can generate preliminary reports based on the ultrasound data, highlighting key findings and suggesting possible diagnoses. This can expedite the review process by healthcare professionals. AI also can help in organizing and managing vast amounts of ultrasound data, making it easier for healthcare professionals to access and retrieve patient information when needed.

Meanwhile, based on the Segment Anything (SA) project [171], Kirillov et al. developed a new segmentation model (SAM). The SAM demonstrates impressive zero-shot performance across various tasks, often matching or exceeding fully supervised methods. This indicates the model’s generalizability and potential applicability to a wide range of segmentation challenges. Numerous studies have adopted the SAM in the medical image segmentation [172,173]. With the continuous development of large language models (LLMs), AI technology based on these models can also be applied to medicine [174–178]. Based on these studies, it can be seen that LLMs can play a significant role in medical ultrasound imaging. It can be used to generate preliminary reports of ultrasound imaging by analyzing the textual descriptions provided by the sonographer or the data obtained from the ultrasound device. The reports will not only save time but also reduce the workload of radiologists. LLMs can generate descriptive annotations for the images based on the features identified through image processing techniques. By working through complex medical language and jargon, LLMs can translate these into more patient-friendly language. This helps patients better understand their medical condition and the significance of their ultrasound results. LLMs can be utilized in creating interactive training material for medical students and professionals. This can assist them in learning the nomenclature, understanding complex medical conditions, and being updated with the latest medical research associated with ultrasound imaging. LLMs can assist in data collection, research conduction, and generating insights from large bodies of medical texts or research papers, offering valuable contributions to the field of medical ultrasound imaging. LLMs can also be integrated with AI and machine learning algorithms aimed at identifying and diagnosing diseases from ultrasound imagery. The LLM can then provide detailed explanations or feedback based on the AI's findings in a way that is understandable for the healthcare provider.

5 Conclusion

In conclusion, the future application of deep learning in medical ultrasound imaging is multifaceted. It can not only enhance image quality and diagnostic efficiency but also promote the development of personalized medicine and precision medicine. With the increasing availability of computational resources and the continuous improvement of algorithms, we can expect deep learning to play an increasingly important role in ultrasound imaging technology.

Author contributions

KS: Funding acquisition, Supervision, Writing–original draft, Writing–review and editing. JF: Conceptualization, Writing–original draft, Writing–review and editing. DC: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Natural Science Foundation Project of CQ CSTC (Grant No. cstc2021jcyj-msxmX0305), by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202102804), by the Research Program of Chongqing University of Education (Grant No. KY201925C), by the Chongqing Electronics Engineering Technology Research Center for Interactive Learning, and by the Chongqing Big Data Engineering Laboratory for Children.

Acknowledgments

We would like to thank the editors and the reviewers for their helpful remarks.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Beutel J. *Handbook of medical imaging*, 3. Bellingham, Washington, United States: Spie Press (2000).

2. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng* (2017) 19:221–48. doi:10.1146/annurev-bioeng-071516-044442

3. Cheplygina V, de Bruijne M, Pluim JP. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med image Anal* (2019) 54:280–96. doi:10.1016/j.media.2019.03.009

4. Wang G, Ye JC, De Man B. Deep learning for tomographic image reconstruction. *Nat machine intelligence* (2020) 2:737–48. doi:10.1038/s42256-020-00273-z

5. Ben Yedder H, Cardoen B, Hamarneh G. Deep learning for biomedical image reconstruction: a survey. *Artif intelligence Rev* (2021) 54:215–51. doi:10.1007/s10462-020-09861-2

6. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE* (2021) 109: 820–38. doi:10.1109/jproc.2021.3054390

7. Koetzier LR, Mastrodicasa D, Szczykutowicz TP, van der Werf NR, Wang AS, Sandfort V, et al. Deep learning image reconstruction for ct: technical principles and clinical prospects. *Radiology* (2023) 306:e221257. doi:10.1148/radiol.221257

8. Kiryu S, Akai H, Yasaka K, Tajima T, Kunimatsu A, Yoshioka N, et al. Clinical impact of deep learning reconstruction in mri. *Radiographics* (2023) 43:e220133. doi:10.1148/rg.220133

9. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health* (2019) 1:e271–97. doi:10.1016/s2589-7500(19)30123-2

10. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. *Radiographics* (2017) 37:2113–31. doi:10.1148/rg.2017170077

11. Bizopoulos P, Koutsouris D. Deep learning in cardiology. *IEEE Rev Biomed Eng* (2018) 12:168–93. doi:10.1109/rbme.2018.2885714

12. Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, et al. Management of thyroid nodules seen on us images: deep learning may match performance of radiologists. *Radiology* (2019) 292:695–701. doi:10.1148/radiol.2019181343

13. Murtaza G, Shuib L, Abdul Wahab AW, Mujtaba G, Mujtaba G, Nweke HF, et al. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artif Intelligence Rev* (2020) 53:1655–720. doi:10.1007/s10462-019-09716-5

14. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med* (2021) 27:775–84. doi:10.1038/s41591-021-01343-4

15. Gul S, Khan MS, Bibi A, Khandakar A, Ayari MA, Chowdhury ME. Deep learning techniques for liver and liver tumor segmentation: a review. *Comput Biol Med* (2022) 147:105620. doi:10.1016/j.compbiomed.2022.105620

16. Zhu Z, Sun M, Qi G, Li Y, Gao X, Liu Y. Sparse dynamic volume transunet with multi-level edge fusion for brain tumor segmentation. *Comput Biol Med* (2024) 172: 108284. doi:10.1016/j.compbiomed.2024.108284

17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* (2019) 25:44–56. doi:10.1038/s41591-018-0300-7

18. Szabo TL. *Diagnostic ultrasound imaging: inside out*. Academic Press (2004).

19. Buchan I, Covvey HD, Rakowski H. An artificial intelligence approach to automatic left ventricular border detection in 2-d echocardiography. In: *Proceedings of the annual symposium on computer application in medical care (American medical informatics association)* (1985). p. 691.

20. Goldberg V, Manduca A, Ewert DL, Gisvold JJ, Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Med Phys* (1992) 19:1475–81. doi:10.1118/1.596804

21. Han W, Birkeland R. Artificial intelligence as an approach to improve ultrasonic log scanning. *Acoust Imaging* (1993) 201–8. doi:10.1007/978-1-4615-2958-3_27

22. Buller D, Buller A, Innocent PR, Pawlak W. Determining and classifying the region of interest in ultrasonic images of the breast using neural networks. *Artif Intelligence Med* (1996) 8:53–66. doi:10.1016/0933-3657(95)00020-8

23. Wu K, Chen X, Ding M. Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. *Optik* (2014) 125:4057–63. doi:10.1016/j.ijleo.2014.01.114

24. Azizi S, Imani F, Zhuang B, Tahmasebi A, Kwak JT, Xu S, et al. Ultrasound-based detection of prostate cancer using automatic feature selection with deep belief networks. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, october 5-9, 2015, proceedings, Part II 18*. Springer (2015). p. 70–7.

25. Shi J, Zhou S, Liu X, Zhang Q, Lu M, Wang T. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing* (2016) 194:87–94. doi:10.1016/j.neucom.2016.01.074

26. Mischi M, Bell MAL, Van Sloun RJ, Eldar YC. Deep learning in medical ultrasound—from image formation to image analysis. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2020) 67:2477–80. doi:10.1109/tuffc.2020.3026598

27. Gasse M, Millioz F, Roux E, Garcia D, Liebgott H, Friboulet D. High-quality plane wave compounding using convolutional neural networks. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2017) 64:1637–9. doi:10.1109/tuffc.2017.2736890

28. Luchies A, Byram B. Deep neural networks for ultrasound beamforming. In: *2017 IEEE international ultrasonics symposium (IUS)*. IEEE (2017). p. 1–4. doi:10.1109/ULTSYM.2017.8092159

29. Montaldo G, Tanter M, Bercoff J, Benech N, Fink M. Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2009) 56:489–506. doi:10.1109/tuffc.2009.1067

30. Luchies AC, Byram BC. Deep neural networks for ultrasound beamforming. *IEEE Trans Med Imaging* (2018) 37:2010–21. doi:10.1109/tmi.2018.2809641

31. Simson W, Paschali M, Navab N, Zahnd G. Deep learning beamforming for sub-sampled ultrasound data. In: *2018 IEEE international ultrasonics symposium (IUS)*. IEEE (2018). p. 1–4.

32. Nair AA, Gubbi MR, Tran TD, Reiter A, Bell MAL. A fully convolutional neural network for beamforming ultrasound images. In: *2018 IEEE international ultrasonics symposium (IUS)*. IEEE (2018). p. 1–4.

33. Nair AA, Tran TD, Reiter A, Bell MAL. A deep learning based alternative to beamforming ultrasound images. In: *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE (2018). p. 3359–63.

34. Nair AA, Tran TD, Reiter A, Bell MAL. One-step deep learning approach to ultrasound image formation and image segmentation with a fully convolutional neural network. In: *2019 IEEE international ultrasonics symposium (IUS)*. IEEE (2019). p. 1481–4.

35. Nair AA, Washington KN, Tran TD, Reiter A, Bell MAL. Deep learning to obtain simultaneous image and segmentation outputs from a single input of raw ultrasound channel data. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2020) 67:2493–509. doi:10.1109/tuffc.2020.2993779

36. Strohm H, Rothlübbers S, Eickel K, Günther M. Deep learning-based reconstruction of ultrasound images from raw channel data. *Int J Comp Assist Radiol Surg* (2020) 15:1487–90. doi:10.1007/s11548-020-02197-w

37. Senouf O, Vedula S, Zurakhov G, Bronstein A, Zibulevsky M, Michailovich O, et al. High frame-rate cardiac ultrasound imaging with deep learning. In: *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, granada, Spain, september 16-20, 2018, proceedings, Part I*. Springer (2018). p. 126–34.

38. Vedula S, Senouf O, Zurakhov G, Bronstein A, Zibulevsky M, Michailovich O, et al. High quality ultrasonic multi-line transmission through deep learning. In: *Machine learning for medical image reconstruction: first international workshop, MLMIR 2018, held in conjunction with MICCAI 2018, granada, Spain, september 16, 2018, proceedings 1*. Springer (2018). p. 147–55.

39. Vedula S, Senouf O, Zurakhov G, Bronstein A, Michailovich O, Zibulevsky M. *Learning beamforming in ultrasound imaging* (2018). arXiv preprint arXiv:1812.08043.

40. Luijten B, Cohen R, De Bruijn FJ, Schmeitz HA, Mischi M, Eldar YC, et al. Deep learning for fast adaptive beamforming. In: *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE (2019). p. 1333–7.

41. Luijten B, Cohen R, De Bruijn FJ, Schmeitz HA, Mischi M, Eldar YC, et al. Adaptive ultrasound beamforming using deep learning. *IEEE Trans Med Imaging* (2020) 39:3967–78. doi:10.1109/tmi.2020.3008537

42. Wiacek A, Gonzalez E, Dehak N, Bell MAL. Coherenet: a deep learning approach to coherence-based beamforming. In: *2019 IEEE international ultrasonics symposium (IUS)*. IEEE (2019). p. 287–90.

43. Wiacek A, González E, Bell MAL. Coherenet: a deep learning architecture for ultrasound spatial correlation estimation and coherence-based beamforming. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2020) 67:2574–83. doi:10.1109/tuffc.2020.2982848

44. Lediju MA, Trahey GE, Byram BC, Dahl JJ. Short-lag spatial coherence of backscattered echoes: imaging characteristics. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2011) 58:1377–88. doi:10.1109/tuffc.2011.1957

45. Yoon YH, Khan S, Huh J, Ye JC. Efficient b-mode ultrasound image reconstruction from sub-sampled rf data using deep learning. *IEEE Trans Med Imaging* (2018) 38:325–36. doi:10.1109/tmi.2018.2864821

46. Mamistvalov A, Eldar YC. Compressed fourier-domain convolutional beamforming for sub-nyquist ultrasound imaging. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2021) 69:489–99. doi:10.1109/tuffc.2021.3123079

47. Mamistvalov A, Amar A, Kessler N, Eldar YC. Deep-learning based adaptive ultrasound imaging from sub-nyquist channel data. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2022) 69:1638–48. doi:10.1109/tuffc.2022.3160859

48. Qi Y, Guo Y, Wang Y. Image quality enhancement using a deep neural network for plane wave medical ultrasound imaging. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2020) 68:926–34. doi:10.1109/tuffc.2020.3023154

49. Chen Y, Liu J, Luo X, Luo J. A self-supervised deep learning approach for high frame rate plane wave beamforming with two-way dynamic focusing. In: *2021 IEEE international ultrasonics symposium (IUS)*. IEEE (2021). p. 1–4.

50. Lu J-Y, Lee P-Y, Huang C-C. Improving image quality for single-angle plane wave ultrasound imaging with convolutional neural network beamformer. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2022) 69:1326–36. doi:10.1109/tuffc.2022.3152689

51. Nguon LS, Seo J, Seo K, Han Y, Park S. Reconstruction for plane-wave ultrasound imaging using modified u-net-based beamformer. *Comput Med Imaging Graphics* (2022) 98:102073. doi:10.1016/j.compmedimag.2022.102073

52. Goudarzi S, Rivaz H. Deep reconstruction of high-quality ultrasound images from raw plane-wave data: a simulation and *in vivo* study. *Ultrasonics* (2022) 125:106778. doi:10.1016/j.ultras.2022.106778

53. Wasih M, Almekkawy M. A robust deep neural network approach for ultrafast ultrasound imaging using single angle plane wave. In: *2022 IEEE international ultrasonics symposium (IUS)*. IEEE (2022). p. 1–4.

54. Seoni S, Salvi M, Matrone G, Meiburger KM. Ultrasound image beamforming optimization using a generative adversarial network. In: *2022 IEEE international ultrasonics symposium (IUS)*. IEEE (2022). p. 1–4.

55. Gao J, Xu L, Zou Q, Zhang B, Wang D, Wan M. A progressively dual reconstruction network for plane wave beamforming with both paired and unpaired training data. *Ultrasonics* (2023) 127:106833. doi:10.1016/j.ultras.2022.106833

56. Mor E, Bar-Hillel A. A unified deep network for beamforming and speckle reduction in plane wave imaging: a simulation study. *Ultrasonics* (2020) 103:106069. doi:10.1016/j.ultras.2020.106069

57. Vienneau E, Luchies A, Byram B. An improved training scheme for deep neural network ultrasound beamforming. In: *2019 IEEE international ultrasonics symposium (IUS)*. IEEE (2019). p. 568–70.

58. Luchies AC, Byram BC. Training improvements for ultrasound beamforming with deep neural networks. *Phys Med Biol* (2019) 64:045018. doi:10.1088/1361-6560/aafd50

59. Tierney J, Luchies A, Berger M, Byram B. Image quality-based regularization for deep network ultrasound beamforming. In: *2020 IEEE international ultrasonics symposium (IUS)*. IEEE (2020). p. 1–3.

60. Bell MAL, Huang J, Hyun D, Eldar YC, Van Sloun R, Mischi M. Challenge on ultrasound beamforming with deep learning (cubdl). In: *2020 IEEE international ultrasonics symposium (IUS)*. IEEE (2020). p. 1–5.

61. Hyun D, Wiacek A, Goudarzi S, Rothlübbers S, Asif A, Eickel K, et al. Deep learning for ultrasound image formation: cubdl evaluation framework and open datasets. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2021) 68:3466–83. doi:10.1109/tuffc.2021.3094849

62. Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, et al. Deep learning in medical ultrasound analysis: a review. *Engineering* (2019) 5:261–75. doi:10.1016/j.eng.2018.11.020

63. Wang Y, Ge X, Ma H, Qi S, Zhang G, Yao Y. Deep learning in medical ultrasound image analysis: a review. *IEEE Access* (2021) 9:54310–24. doi:10.1109/access.2021.3071301

64. Fujioka T, Mori M, Kubota K, Oyama J, Yamaga E, Yashima Y, et al. The utility of deep learning in breast ultrasonic imaging: a review. *Diagnostics* (2020) 10:1055. doi:10.3390/diagnostics10121055

65. Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol* (2018) 91:20170576. doi:10.1259/bjr.20170576

66. Xu Y, Wang Y, Yuan J, Cheng Q, Wang X, Carson PL. Medical breast ultrasound image segmentation by machine learning. *Ultrasonics* (2019) 91:1–9. doi:10.1016/j.ultras.2018.07.006

67. Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat Biomed Eng* (2021) 5:522–32. doi:10.1038/s41551-021-00711-2

68. Chen C, Wang Y, Niu J, Liu X, Li Q, Gong X. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. *IEEE Trans Med Imaging* (2021) 40:2439–51. doi:10.1109/tmi.2021.3078370

69. Jabeen K, Khan MA, Alhaisoni M, Tariq U, Zhang Y-D, Hamza A, et al. Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors* (2022) 22:807. doi:10.3390/s22030807

70. Raza A, Ullah N, Khan JA, Assam M, Guzzo A, Aljuaid H. Deepbreastcancernet: a novel deep learning model for breast cancer detection using ultrasound images. *Appl Sci* (2023) 13:2082. doi:10.3390/app13042082

71. Degel MA, Navab N, Albarqouni S. Domain and geometry agnostic cnns for left atrium segmentation in 3d ultrasound. In: *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, granada, Spain, september 16-20, 2018, proceedings, Part IV 11*. Springer (2018). p. 630–7.

72. Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Trans Med Imaging* (2019) 38:2198–210. doi:10.1109/tmi.2019.2900516

73. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, et al. Deep learning interpretation of echocardiograms. *NPJ digital Med* (2020) 3:10. doi:10.1038/s41746-019-0216-8

74. Narang A, Bae R, Hong H, Thomas Y, Surette S, Cadieu C, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA Cardiol* (2021) 6:624–32. doi:10.1001/jamacardio.2021.0185

75. Ha EJ, Baek JH. Applications of machine learning and deep learning to thyroid imaging: where do we stand? *Ultrasonography* (2021) 40:23–9. doi:10.14366/usg.20068

76. Park VY, Han K, Seong YK, Park MH, Kim E-K, Moon HJ, et al. Diagnosis of thyroid nodules: performance of a deep learning convolutional neural network model vs. radiologists. *Scientific Rep* (2019) 9:17843. doi:10.1038/s41598-019-54434-1

77. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *The Lancet Digital Health* (2021) 3:e250–9. doi:10.1016/s2589-7500(21)00041-8

78. Buda M, Wildman-Tobriner B, Castor K, Hoang JK, Mazurowski MA. Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images. *Ultrasound Med Biol* (2020) 46:415–21. doi:10.1016/j.ultrasmedbio.2019.10.003

79. Wang L, Zhang L, Zhu M, Qi X, Yi Z. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. *Med image Anal* (2020) 61:101665. doi:10.1016/j.media.2020.101665

80. Zhao Z, Yang C, Wang Q, Zhang H, Shi L, Zhang Z. A deep learning-based method for detecting and classifying the ultrasound images of suspicious thyroid nodules. *Med Phys* (2021) 48:7959–70. doi:10.1002/mp.15319

81. Zhu Y-C, AlZoubi A, Jassim S, Jiang Q, Zhang Y, Wang Y-B, et al. A generic deep learning framework to classify thyroid and breast lesions in ultrasound images. *Ultrasonics* (2021) 110:106300. doi:10.1016/j.ultras.2020.106300

82. Feng Y, Yang F, Zhou X, Guo Y, Tang F, Ren F, et al. A deep learning approach for targeted contrast-enhanced ultrasound based prostate cancer detection. *IEEE/ACM Trans Comput Biol Bioinformatics* (2018) 16:1794–801. doi:10.1109/tcbb.2018.2835444

83. Azizi S, Bayat S, Yan P, Tahmasebi A, Kwak JT, Xu S, et al. Deep recurrent neural networks for prostate cancer detection: analysis of temporal enhanced ultrasound. *IEEE Trans Med Imaging* (2018) 37:2695–703. doi:10.1109/tmi.2018.2849959

84. Hassan MR, Islam MF, Uddin MZ, Ghoshal G, Hassan MM, Huda S, et al. Prostate cancer classification from ultrasound and mri images using deep learning based explainable artificial intelligence. *Future Generation Comp Syst* (2022) 127:462–72. doi:10.1016/j.future.2021.09.030

85. Wang Y, Deng Z, Hu X, Zhu L, Yang X, Xu X, et al. Deep attentional features for prostate segmentation in ultrasound. In: *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, granada, Spain, september 16-20, 2018, proceedings, Part IV 11*. Springer (2018). p. 523–30.

86. Anas EMA, Mousavi P, Abolmaesumi P. A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy. *Med image Anal* (2018) 48:107–16. doi:10.1016/j.media.2018.05.010

87. Lei Y, Tian S, He X, Wang T, Wang B, Patel P, et al. Ultrasound prostate segmentation based on multidirectional deeply supervised v-net. *Med Phys* (2019) 46:3194–206. doi:10.1002/mp.13577

88. Wang Y, Dou H, Hu X, Zhu L, Yang X, Xu M, et al. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. *IEEE Trans Med Imaging* (2019) 38:2768–78. doi:10.1109/tmi.2019.2913184

89. Karimi D, Zeng Q, Mathur P, Avinash A, Mahdavi S, Spadinger I, et al. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Med image Anal* (2019) 57:186–96. doi:10.1016/j.media.2019.07.005

90. Orlando N, Gillies DJ, Gyacskov I, Romagnoli C, D'Souza D, Fenster A. Automatic prostate segmentation using deep learning on clinically diverse 3d transrectal ultrasound images. *Med Phys* (2020) 47:2413–26. doi:10.1002/mp.14134

91. Orlando N, Gyacskov I, Gillies DJ, Guo F, Romagnoli C, D'Souza D, et al. Effect of dataset size, image quality, and image type on deep learning-based automatic prostate segmentation in 3d ultrasound. *Phys Med Biol* (2022) 67:074002. doi:10.1088/1361-6560/ac5a93

92. Fiorentino MC, Villani FP, Di Cosmo M, Frontoni E, Moccia S. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Med image Anal* (2023) 83:102629. doi:10.1016/j.media.2022.102629

93. Sobhaninia Z, Rafiei S, Emami A, Karimi N, Najarian K, Samavi S, et al. Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. In: *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE (2019). p. 6545–8.

94. van den Heuvel TL, Petros H, Santini S, de Korte CL, van Ginneken B. Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries. *Ultrasound Med Biol* (2019) 45:773–85. doi:10.1016/j.ultrasmedbio.2018.09.015

95. Xie H, Wang N, He M, Zhang L, Cai H, Xian J, et al. Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal. *Ultrasound Obstet Gynecol* (2020) 56:579–87. doi:10.1002/uog.21967

96. Pu B, Li K, Li S, Zhu N. Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. *IEEE Trans Ind Inform* (2021) 17:7771–80. doi:10.1109/tii.2021.3069470

97. Komatsu M, Sakai A, Komatsu R, Matsuoka R, Yasutomi S, Shozu K, et al. Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning. *Appl Sci* (2021) 11:371. doi:10.3390/app11010371

98. Milletari F, Ahmadi S-A, Kroll C, Plate A, Rozanski V, Maiostre J, et al. Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound. *Computer Vis Image Understanding* (2017) 164:92–102. doi:10.1016/j.cviu.2017.04.002

99. Ilunga-Mbuyamba E, Avina-Cervantes JG, Lindner D, Arlt F, Ituna-Yudonago JF, Chalopin C. Patient-specific model-based segmentation of brain tumors in 3d intraoperative ultrasound images. *Int J Comput Assist Radiol Surg* (2018) 13:331–42. doi:10.1007/s11548-018-1703-0

100. Xie B, Lei T, Wang N, Cai H, Xian J, He M, et al. Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. *Int J Comp Assist Radiol Surg* (2020) 15:1303–12. doi:10.1007/s11548-020-02182-3

101. Hesse LS, Aliasi M, Moser F, Haak MC, Xie W, Jenkinson M, et al. Subcortical segmentation of the fetal brain in 3d ultrasound using deep learning. *NeuroImage* (2022) 254:119117. doi:10.1016/j.neuroimage.2022.119117

102. Di Ianni T, Airan RD. Deep-fus: a deep learning platform for functional ultrasound imaging of the brain using sparse data. *IEEE Trans Med Imaging* (2022) 41:1813–25. doi:10.1109/tmi.2022.3148728

103. Soleimani M, Rymarczyk T, Kłosowski G. Ultrasound brain tomography: comparison of deep learning and deterministic methods. *IEEE Trans Instrumentation Meas* (2023) 73:1–12. doi:10.1109/tim.2023.3330229

104. Van Sloun RJ, Cohen R, Eldar YC. Deep learning in ultrasound imaging. *Proc IEEE* (2019) 108:11–29. doi:10.1109/jproc.2019.2932116

105. van Sloun RJ, Ye JC, Eldar YC. *1 deep learning for ultrasound beamforming* (2021). arXiv preprint arXiv:2109.11431.

106. Luijten B, Chennakeshava N, Eldar YC, Mischi M, van Sloun RJ. Ultrasound signal processing: from models to deep learning. *Ultrasound Med Biol* (2023) 49:677–98. doi:10.1016/j.ultrasmedbio.2022.11.003

107. Afrin H, Larson NB, Fatemi M, Alizad A. Deep learning in different ultrasound methods for breast cancer, from diagnosis to prognosis: current trends, challenges, and an analysis. *Cancers* (2023) 15:3139. doi:10.3390/cancers15123139

108. Akkus Z, Aly YH, Attia IZ, Lopez-Jimenez F, Arruda-Olson AM, Pellikka PA, et al. Artificial intelligence (ai)-empowered echocardiography interpretation: a state-of-the-art review. *J Clin Med* (2021) 10:1391. doi:10.3390/jcm10071391

109. Khachnaoui H, Guetari R, Khlifa N. A review on deep learning in thyroid ultrasound computer-assisted diagnosis systems. In: *2018 IEEE international conference on image processing, applications and systems (IPAS)*. IEEE (2018). p. 291–7.

110. Ali M, Magee D, Dasgupta U. *Signal processing overview of ultrasound systems for medical imaging*. Texas: SPRAB12, Texas Instruments (2008). p. 55.

111. Thomenius KE. Evolution of ultrasound beamformers. In: *1996 IEEE ultrasonics symposium. Proceedings (IEEE)*, 2. IEEE (1996). p. 1615–22. doi:10.1109/ultsym.1996.584398

112. Perrot V, Polichetti M, Varray F, Garcia D. So you think you can das? a viewpoint on delay-and-sum beamforming. *Ultrasonics* (2021) 111:106309. doi:10.1016/j.ultras.2020.106309

113. Synnevag JF, Austeng A, Holm S. Adaptive beamforming applied to medical ultrasound imaging. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2007) 54: 1606–13. doi:10.1109/tuffc.2007.431

114. Ortiz SHC, Chiu T, Fox MD. Ultrasound image enhancement: a review. *Biomed Signal Process Control* (2012) 7:419–28. doi:10.1016/j.bspc.2012.02.002

115. Basset O, Cachard C. Ultrasound image post-processing–application to segmentation. In: *Physics for medical imaging applications*. Springer (2007). p. 227–39.

116. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans image Process* (2004) 13:600–12. doi:10.1109/tip.2003.819861

117. Long J, Trahey G, Bottenus N. Spatial coherence in medical ultrasound: a review. *Ultrasound Med Biol* (2022) 48:975–96. doi:10.1016/j.ultrasmedbio.2022.01.009

118. Hollman K, Rigby K, O'donnell M. Coherence factor of speckle from a multi-row probe. In: *1999 IEEE ultrasonics symposium. Proceedings. International symposium (cat. No. 99CH37027)*, 2. IEEE (1999). p. 1257–60. doi:10.1109/ultsym.1999.849225

119. Li P-C, Li M-L. Adaptive imaging using the generalized coherence factor. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2003) 50:128–41. doi:10.1109/tuffc.2003.1182117

120. Camacho J, Parrilla M, Fritsch C. Phase coherence imaging. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2009) 56:958–74. doi:10.1109/tuffc.2009.1128

121. Fornage BD. Ultrasound of the breast. *Ultrasound Q* (1993) 11:1–40. doi:10.1097/00013644-199300000-00001

122. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data in brief* (2020) 28:104863. doi:10.1016/j.dib.2019.104863

123. Shao F, Ling KV, Ng WS, Wu RY. Prostate boundary detection from ultrasonographic images. *J Ultrasound Med* (2003) 22:605–23. doi:10.7863/jum.2003.22.6.605

124. Xu L, Gao S, Shi L, Wei B, Liu X, Zhang J, et al. Exploiting vector attention and context prior for ultrasound image segmentation. *Neurocomputing* (2021) 454:461–73. doi:10.1016/j.neucom.2021.05.033

125. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer (2015). p. 234–41.

126. Ravichandran G, Fellows G. The accuracy of a hand-held real time ultrasound scanner for estimating bladder volume. *Br J Urol* (1983) 55:25–7. doi:10.1111/j.1464-410x.1983.tb07073.x

127. Ireton RC, Krieger JN, Cardenas DD, Williams-Burden B, Kelly E, Souci T, et al. Bladder volume determination using a dedicated, portable ultrasound scanner. *J Urol* (1990) 143:909–11. doi:10.1016/s0022-5347(17)40133-9

128. Coombes GM, Millard RJ. The accuracy of portable ultrasound scanning in the measurement of residual urine volume. *J Urol* (1994) 152:2083–5. doi:10.1016/s0022-5347(17)32314-5

129. Teng C-H, Huang Y-H, Kuo B-J, Bih L-I. Application of portable ultrasound scanners in the measurement of post-void residual urine. *J Nurs Res* (2005) 13:216–24. doi:10.1097/01.jnr.0000387543.68383.a0

130. Luo H, Jin F, Yang D, Wang Y, Li C, Guo M, et al. Interfractional variation in bladder volume and its impact on cervical cancer radiotherapy: clinical significance of portable bladder scanner. *Med Phys* (2016) 43:4412–9. doi:10.1118/1.4954206

131. Zhao L, Liao L, Gao L, Gao Y, Chen G, Cong H, et al. Effects of bladder shape on accuracy of measurement of bladder volume using portable ultrasound scanner and development of correction method. *Neurourology and Urodynamics* (2019) 38:653–9. doi:10.1002/nau.23883

132. Ohira S, Komiyama R, Kanayama N, Sakai K, Hirata T, Yoshikata K, et al. Improvement in bladder volume reproducibility using a-mode portable ultrasound bladder scanner in moderate-hypofractionated volumetric modulated arc therapy for prostate cancer patients. *J Appl Clin Med Phys* (2022) 23:e13546. doi:10.1002/acm2.13546

133. Jeong E, Bae S, Park M, Jung W, Kang J, Song T-K. Color Doppler imaging on a smartphone-based portable us system: preliminary study. In: *2015 IEEE international ultrasonics symposium (IUS)*. IEEE (2015). p. 1–4.

134. Di Ianni T, Hoyos CAV, Ewertsen C, Kjeldsen TK, Mosegaard J, Nielsen MB, et al. A vector flow imaging method for portable ultrasound using synthetic aperture sequential beamforming. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2017) 64:1655–65. doi:10.1109/tuffc.2017.2742599

135. Jafari MH, Girgis H, Van Woudenberg N, Moulson N, Luong C, Fung A, et al. Cardiac point-of-care to cart-based ultrasound translation using constrained cyclegan. *Int J Comput Assist Radiol Surg* (2020) 15:877–86. doi:10.1007/s11548-020-02141-y

136. Seviaryn F, Malyarenko E, Schreiner G, Seviaryna I, Maev RG. Handheld high-resolution ultrasonic scanner for quantitative assessment of skin conditions. In: *2019 IEEE international ultrasonics symposium (IUS)*. IEEE (2019). p. 2380–2.

137. Qian F, Zhou X, Zhou J, Liu Z, Nie Q. A valuable and affordable handheld ultrasound in combating covid-19. *Crit Care* (2020) 24:334–2. doi:10.1186/s13054-020-03064-5

138. Bennett D, De Vita E, Mezzasalma F, Lanzarone N, Cameli P, Bianchi F, et al. Portable pocket-sized ultrasound scanner for the evaluation of lung involvement in coronavirus disease 2019 patients. *Ultrasound Med Biol* (2021) 47:19–24. doi:10.1016/j.ultrasmedbio.2020.09.014

139. Aminlari A, Quenzer F, Hayden S, Stone J, Murchison C, Campbell C. A case of covid-19 diagnosed at home with portable ultrasound and confirmed with home serology test. *J Emerg Med* (2021) 60:399–401. doi:10.1016/j.jemermed.2020.10.022

140. Baran JM, Webster JG. Design of low-cost portable ultrasound systems. In: *2009 annual international conference of the IEEE engineering in medicine and biology society*. IEEE (2009). p. 792–5.

141. Xu X, Venkataraman H, Oswal S, Bartolome E, Vasanth K. Challenges and considerations of analog front-ends design for portable ultrasound systems. In: *2010 IEEE international ultrasonics symposium (IEEE)*. IEEE (2010). p. 310–3.

142. Kim G-D, Yoon C, Kye S-B, Lee Y, Kang J, Yoo Y, et al. A single fpga-based portable ultrasound imaging system for point-of-care applications. *IEEE Trans Ultrason ferroelectrics, frequency Control* (2012) 59:1386–94. doi:10.1109/tuffc.2012.2339

143. Kang J, Yoon C, Lee J, Kye S-B, Lee Y, Chang JH, et al. A system-on-chip solution for point-of-care ultrasound imaging systems: architecture and asic implementation. *IEEE Trans Biomed circuits Syst* (2015) 10:412–23. doi:10.1109/tbcas.2015.2431272

144. Donoho DL. Compressed sensing. *IEEE Trans Inf Theor* (2006) 52:1289–306. doi:10.1109/tit.2006.871582

145. Zhou J, Hoyos S, Sadler BM. Asynchronous compressed beamformer for portable diagnostic ultrasound systems. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* (2014) 61:1791–801. doi:10.1109/tuffc.2014.006384

146. Shin B, Jeon S, Ryu J, Kwon HJ. Compressed sensing for elastography in portable ultrasound. *Ultrason Imaging* (2017) 39:393–413. doi:10.1177/0161734617716938

147. George SS, Mitrovic J, Anand A, Ignjatovic Z. Low-complexity compressive beamforming for portable ultrasound imaging. In: *2017 IEEE international ultrasonics symposium (IUS)*. IEEE (2017). p. 1–4.

148. Mitrovic J, La Pietra L, Ignjatovic Z. Portable ultrasound through compressive beamforming with improved contrast. In: *2018 IEEE international ultrasonics symposium (IUS)*. IEEE (2018). p. 1–4.

149. Zhou Z, Wang Y, Guo Y, Qi Y, Yu J. Image quality improvement of hand-held ultrasound devices with a two-stage generative adversarial network. *IEEE Trans Biomed Eng* (2019) 67:298–311. doi:10.1109/tbme.2019.2912986

150. Zhou Z, Siddiquee MR, Tajbakhsh N, Liang J (2018). Unet++: a nested u-net architecture for medical image segmentation. In: *Deep learn med image anal multimodal learn clin decis support (2018), deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018*. Granada, Spain: Springer, Cham 11045, 3–11. doi:10.1007/978-3-030-00889-5_1

151. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: *Medical image computing and computer-assisted intervention–MICCAI 2016: 19th international conference, athens, Greece, october 17-21, 2016, proceedings, Part II 19*. Springer (2016). p. 424–32.

152. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. *Attention u-net: learning where to look for the pancreas* (2018). *arXiv preprint arXiv:1804.03999*.

153. Ibtehaz N, Rahman MS. Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks* (2020) 121:74–87. doi:10.1016/j.neunet.2019.08.025

154. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30. doi:10.5555/3295222.3295349

155. Gao Y, Zhou M, Metaxas DN. Utnet: a hybrid transformer architecture for medical image segmentation. In: *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, strasbourg, France, september 27–october 1, 2021, proceedings, Part III 24*. Springer (2021). p. 61–71.

156. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. *Transunet: Transformers make strong encoders for medical image segmentation* (2021). arXiv preprint arXiv:2102.04306.

157. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3d tumor segmentation. In: *International conference on medical image computing and computer-assisted intervention*. Springer (2022). p. 162–72.

158. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: *European conference on computer vision*. Springer (2022). p. 205–18.

159. Shi L, Gao T, Zhang Z, Zhang J. Stm-unet: an efficient u-shaped architecture based on swin transformer and multiscale mlp for medical image segmentation. In: GLOBECOM 2023-2023 IEEE global communications conference *(IEEE)*. IEEE (2023). p. 2003–8.

160. Yang S, Niu J, Wu J, Wang Y, Liu X, Li Q. Automatic ultrasound image report generation with adaptive multimodal attention mechanism. *Neurocomputing* (2021) 427:40–9. doi:10.1016/j.neucom.2020.09.084

161. Liang J, Yang X, Huang Y, Li H, He S, Hu X, et al. Sketch guided and progressive growing gan for realistic and editable ultrasound image synthesis. *Med image Anal* (2022) 79:102461. doi:10.1016/j.media.2022.102461

162. Li C, Wang L, Li Y. Transformer and group parallel axial attention co-encoder for medical image segmentation. *Scientific Rep* (2022) 12:16117. doi:10.1038/s41598-022-20440-z

163. Xu S, Quan H. Ect-nas: searching efficient cnn-transformers architecture for medical image segmentation. In: *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE (2021). p. 1601–4.

164. Zhou H-Y, Guo J, Zhang Y, Yu L, Wang L, Yu Y. *nnformer: interleaved transformer for volumetric segmentation* (2021). arXiv preprint arXiv:2109.03201.

165. Liu D, Gao Y, Zhangli Q, Han L, He X, Xia Z, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In: *International conference on medical image computing and computer-assisted intervention*. Springer (2022). p. 485–95.

166. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. *Inf Fusion* (2023) 91: 376–87. doi:10.1016/j.inffus.2022.10.022

167. Liu Y, Yu C, Cheng J, Wang ZJ, Chen X. Mm-net: a mixformer-based multi-scale network for anatomical and functional image fusion. *IEEE Trans Image Process a Publ IEEE Signal Process Soc* (2024) 33:2197–212. doi:10.1109/tip.2024.3374072

168. Chi J, Li Z, Sun Z, Yu X, Wang H. Hybrid transformer unet for thyroid segmentation from ultrasound scans. *Comput Biol Med* (2023) 153:106453. doi:10.1016/j.compbiomed.2022.106453

169. Wang Y, Ye Z, Wen M, Liang H, Zhang X. Transvfs: a spatio-temporal local-global transformer for vision-based force sensing during ultrasound-guided prostate biopsy. *Med Image Anal* (2024) 94:103130. doi:10.1016/j.media.2024.103130

170. Ahmadi N, Tsang M, Gu A, Tsang T, Abolmaesumi P. Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series. *IEEE Trans Med Imaging* (2024) 43:366–76. doi:10.1109/tmi.2023.3305384

171. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE (2023). p. 4015–26.

172. Lin X, Xiang Y, Zhang L, Yang X, Yan Z, Yu L. *Samus: adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation* (2023). arXiv preprint arXiv:2309.06824.

173. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun* (2024) 15:654. doi:10.1038/s41467-024-44824-z

174. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* (2023) 620:172–80. doi:10.1038/s41586-023-06291-2

175. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* (2023) 29:1930–40. doi:10.1038/s41591-023-02448-8

176. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. *Towards expert-level medical question answering with large language models* (2023). arXiv preprint arXiv:2305.09617.

177. Wu S-H, Tong W-J, Li M-D, Hu H-T, Lu X-Z, Huang Z-R, et al. Collaborative enhancement of consistency and accuracy in us diagnosis of thyroid nodules using large language models. *Radiology* (2024) 310:e232255. doi:10.1148/radiol.232255

178. Sultan LR, Mohamed MK, Andronikou S. *Chatgpt-4: a breakthrough in ultrasound image analysis* (2024).

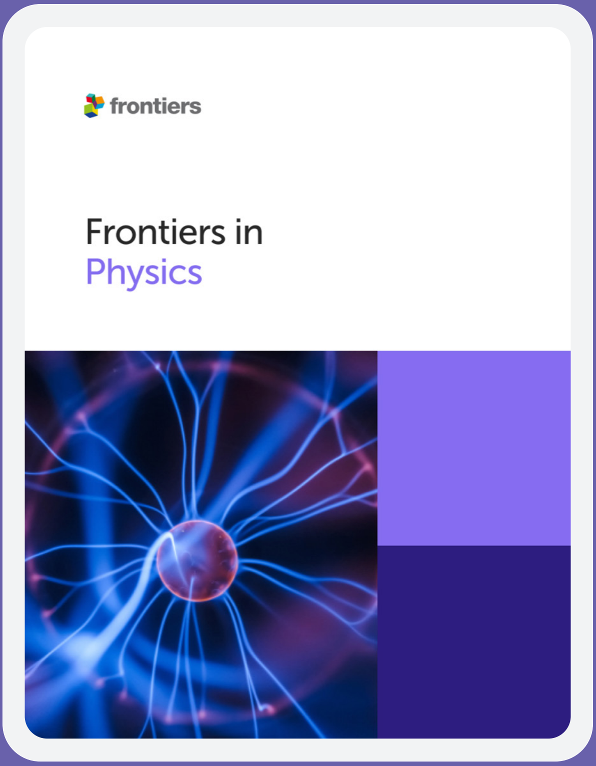
Frontiers in Physics

Investigates complex questions in physics to understand the nature of the physical world

Addresses the biggest questions in physics, from macro to micro, and from theoretical to experimental and applied physics.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Physics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact