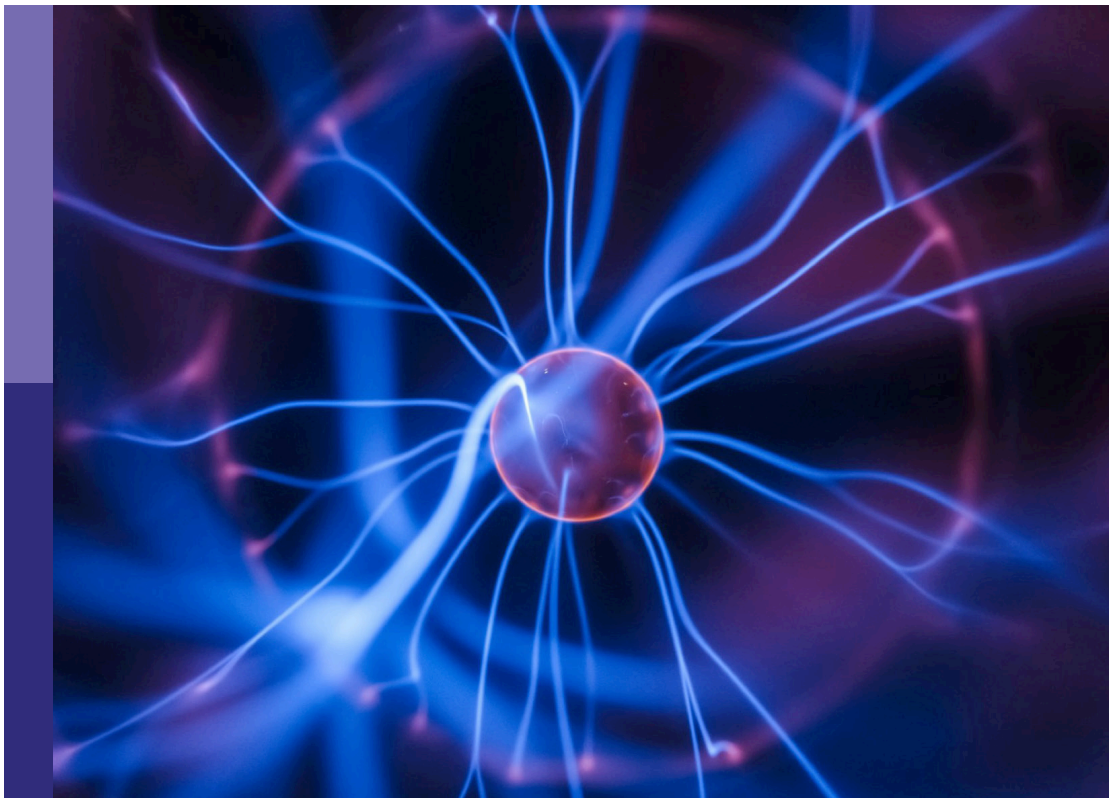
frontiers | Research Topics

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

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

Published in
Frontiers in Physics



ISSN 1664-8714
ISBN 978-2-8325-3693-3
DOI 10.3389/978-2-8325-3693-3

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

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

Citation

Zhu, Z., Liu, Y., Li, H., Qi, G., Xiao, B., eds. (2023). *Multi-sensor imaging and fusion: Methods, evaluations, and applications*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3693-3

Table of contents

frontiers | Frontiers in Physics

TYPE Editorial
PUBLISHED 28 September 2023
DOI 10.3389/fphy.2023.1297201

OPEN ACCESS

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

*CORRESPONDENCE
Guanqiu Qi,
qig@buffalostate.edu

RECEIVED 19 September 2023
ACCEPTED 21 September 2023
PUBLISHED 28 September 2023

CITATION
Qi G, Zhu Z, Liu Y, Li H and Xiao B (2023), Editorial: Multi-sensor imaging and fusion: methods, evaluations, and applications.
Front. Phys. 11:1297201.
doi: 10.3389/fphy.2023.1297201

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

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

[1]Computer Information Systems Department, State University of New York at Buffalo State, 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

KEYWORDS

image processing, image fusion (IF), deep learning—artificial intelligence, multi-sensor image fusion, machine learning

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

Introduction

The technology of multi-sensor imaging and fusion plays an increasingly important role in various fields such as remote sensing [1], medical imaging [2], contraband detection [3], and engineering construction [4]. Multi-sensor image fusion focuses on processing images of the same object or scene captured by multiple sensors, which complement and combine various sensors with multi-level and multi-spatial information, ultimately providing a consistent interpretation of the observed environment [5]. In recent years, multi-sensor image fusion has become a highly active topic, and various fusion methods have been proposed. Moreover, the performance evaluation and downstream applications of multi-sensor imaging and fusion technology are receiving increasing attention. This Research Topic highlights advanced research related to multi-sensor imaging and fusion technology, including image detection and fusion methods, objective evaluation methods, and specific applications in engineering problems. After a thorough peer-review process, all 17 of the articles submitted to this Research Topic were accepted for publication. The following summarizes the main research findings of these works from three aspects.

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

Object detection is an important application of multi-sensor imaging and fusion technologies. He et al. proposed a deep learning object detection network, MSS-YOLOv5, which integrates multi-scale information to enhance feature robustness, improves pooling methods to capture more details, and introduces an angle cost with new weights to accelerate network convergence and improve accuracy. Yang et al. proposed a multi spectral pedestrian detection algorithm that includes a cascaded information

enhancement module and a cross-modal attention feature fusion module to enhance pedestrian features and reduce background interference. Chen et al. proposed a network guided by atomic number Z (ZPGNet), which is used to accurately detect prohibited items in complex X-ray images while reducing the collection cost of atomic number images. Zhou et al. proposed an unsupervised smoke detection algorithm that reduces domain differences and improves the generalization ability of the model through feature alignment and fusion. Meanwhile, multi-level feature fusion of network depth enhances the recognition ability of small targets. Wang et al. proposed an unsupervised method that uses an asymmetric convolution feature extraction network and a pose estimation network with attention mechanisms to solve the problem of monocular depth estimation. They also used a loss function that minimizes the reprojection error to solve the occlusion problem in the projection process. Chen et al. proposed a two-stage domain gap-aware framework to eliminate the bias between the synthetic low-light and real low-light domains, thereby enhancing the generalization capability of low-light image enhancement methods. By utilizing a reverse domain distance guidance strategy, the network can better handle low-light image areas that do not align with the real-world distribution. Liu et al. proposed a new method for the fusion of infrared and visible light images, optimizing edge detail through separate processing of source images and edge detail information. Their two-branch framework extracts features and edge map features directly from source images, and a large number of experiments have verified the effectiveness of their method. Zhou et al. proposed a model to solve the problem of semantic alignment and feature extraction in person text-image matching. The model achieves more efficient feature matching and extraction by adding consistent, clear semantic information and applying an information supplementation network. Li et al. proposed a novel cross-modal hashing method named FSSPDH, which preserves the intrinsic attributes of each modality by learning the hash codes of each modality and constructing a fine-grained similarity matrix. In addition, they used quantization loss to learn hash codes, effectively reducing information loss during the quantization process. Jin et al. proposed a polarization image fusion method that fuses intensity images and linear polarization degrees. It processes the base layer and detail layer through quality evaluation and attention mechanisms. The base layer ensures high contrast of the fused image through a quality evaluation unit, and the detail layer improves the preservation of detail information through an attention enhancement unit.

Objective evaluation methods in multi-sensor imaging

In medical image analysis and evaluation, Xu et al. developed a 16-electrode capacitance imaging (ECT) system for two-dimensional tomography of intracerebral hemorrhage (ICH). The feasibility of ECT in ICH imaging was confirmed through simulation and physical experiments. Qiu et al. retrospectively evaluated patients who underwent stent-assisted coiling (SAC) for intracranial aneurysms, focusing primarily on the rate of embolization and complications. The results showed that all hemodynamic parameters significantly decreased after SAC with four different stents, and laser-cut stents seemed to be more effective than woven stents in reducing aneurysm hemodynamics. Finally, there was no significant difference between the follow-up RROC grades of the four stents. In traffic safety evaluation, Yang et al. proposed a method for immersive tunnel traffic safety evaluation based on the degradation of lighting performance using big data technology. The method utilized numerical simulation, small target recognition tests, and developed a real-time model to illustrate the relationship between the degradation of lighting performance and visual cognition.

Specific applications of multi-sensor technology in engineering problems

Multi-sensor technologies play a significant role in fault detection and signal monitoring. Liu et al. proposed a novel method that combines Improved Energy Fluctuation Index (IEFI) and Modified Variational Mode Decomposition (MVMD) to overcome limitations related to the mode number and balancing parameters. This method can effectively resist interference and accurately extract fault features. Experimental results demonstrate its superior performance in fault signal detection. Meanwhile, Guo et al., by introducing close-range photo grammetry, successfully monitored the differential deformation of immersed tunnel element joints. They not only developed a micro-displacement correction algorithm based on three-dimensional calibration objects, but also a fully automatic system for monitoring the differential deformation of immersed tunnel element joints. In emotion analysis, Yan et al. proposed a Modal Smoothing Fusion Network (MSFNet) that can effectively bridge the semantic gap between text and image at the aspect level of emotional expression. Through feature smoothing and multi-channel attention mechanisms, the model has improved performance in emotion classification. Facing the challenge of defect classification, Liang et al. proposed a multi-level semantic method based on residual adversarial learning for sample enhancement and defect classification. By introducing residual modules and multiple convolutional layers, the network structure is optimized, and the feature extraction capability is enhanced. A multi-level semantic extractor is designed, combined with Wasserstein loss, to solve the instability of network training. This method can generate high-quality defect samples and accurately classify defects.

Conclusion

To conclude, a wide range of related topics have been collected for the special issue. Especially some of the hot Research Topics are from object detection, medical image analysis and evaluation, signal monitoring and fault detection.

Special thanks to Frontier in Physics for the support and efforts provided to this special issue. We would also like to thank all the

authors who contributed their original work to this special issue and all the reviewers for sharing their thoughts on the submissions. We hope that this special issue can inspire the researchers in the 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.

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

References

1. Zhu Z, Luo Y, Qi G, Meng J, Li Y, Mazur N. Remote sensing image defogging networks based on dual self-attention boost residual octave convolution. *Remote Sensing* (2021) 13(16):3104. doi:10.3390/rs13163104

2. 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

3. Zhu Z, Qi G, Lei Y, Jiang D, Mazur N, Liu Y, et al. A long short-term memory neural network based simultaneous quantitative analysis of multiple tobacco chemical components by near-infrared hyperspectroscopy images. *Chemosensors* (2022) 10(5): 164. doi:10.3390/chemosensors10050164

4. Tang L, Huang H, Zhang Y, Qi G, Yu Z. Structure-embedded ghosting artifact suppression network for high dynamic range image reconstruction. *Knowledge-Based Syst* (2023) 263:110278. doi:10.1016/j.knosys.2023.110278

5. 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 Instrumentation Meas* (2021) 70:1–23. doi:10.1109/tim.2020.3024335

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fphy.2022.1117261

OPEN ACCESS

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

REVIEWED BY
Lulu Wang,
Kunming University of Science and Technology, China
Neng Dong,
Nanjing University of Science and Technology, China
Xiaosong Li,
Foshan University, China

*CORRESPONDENCE
Jiaxu Leng,
lengjx@cqupt.edu.cn

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 06 December 2022
ACCEPTED 19 December 2022
PUBLISHED 05 January 2023

CITATION
Chen J, Leng J, Gao X, Mo M and Guan S (2023), Atomic number prior guided network for prohibited items detection from heavily cluttered X-ray imagery. *Front. Phys.* 10:1117261. doi: 10.3389/fphy.2022.1117261

Atomic number prior guided network for prohibited items detection from heavily cluttered X-ray imagery

Jinwen Chen, Jiaxu Leng*, Xinbo Gao, Mengjingcheng Mo and Shibo Guan

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

Prohibited item detection in X-ray images is an effective measure to maintain public safety. Recent prohibited item detection methods based on deep learning has achieved impressive performance. Some methods improve prohibited item detection performance by introducing prior knowledge of prohibited items, such as the edge and size of an object. However, items within baggage are often placed randomly, resulting in cluttered X-ray images, which can seriously affect the correctness and effectiveness of prior knowledge. In particular, we find that different material items in X-ray images have clear distinctions according to their atomic number Z information, which is vital to suppress the interference of irrelevant background information by mining material cues. Inspired by this observation, in this paper, we combined the atomic number Z feature and proposed a novel atomic number Z Prior Guided Network (ZPGNet) to detect prohibited objects from heavily cluttered X-ray images. Specifically, we propose a Material Activation (MA) module that cross-scale flows the atomic number Z information through the network to mine material clues and reduce irrelevant information interference in detecting prohibited items. However, collecting atomic number images requires much labor, increasing costs. Therefore, we propose a method to automatically generate atomic number Z images by exploring the color information of X-ray images, which significantly reduces the manual acquisition cost. Extensive experiments demonstrate that our method can accurately and robustly detect prohibited items from heavily cluttered X-ray images. Furthermore, we extensively evaluate our method on HiXray and OPIXray, and the best result is 2.1% mAP_{50} higher than the state-of-the-art models on HiXray.

KEYWORDS

object detection, X-ray image, prohibited items detection, prior knowledge, public safety

1 Introduction

As society develops, the flow of people on public transport is increasing. X-ray security machine is widely used in the security inspection of railway stations and airports, which is a critical facility for maintaining public safety and transportation safety. However, traditional security checks mostly rely on manual identification methods. After prolonged work hours, security inspectors easily cause fatigue, significantly increasing the risk of missed and false detection and laying many hidden dangers for public safety. Therefore, it is increasingly necessary to identify prohibited items through intelligent algorithms.

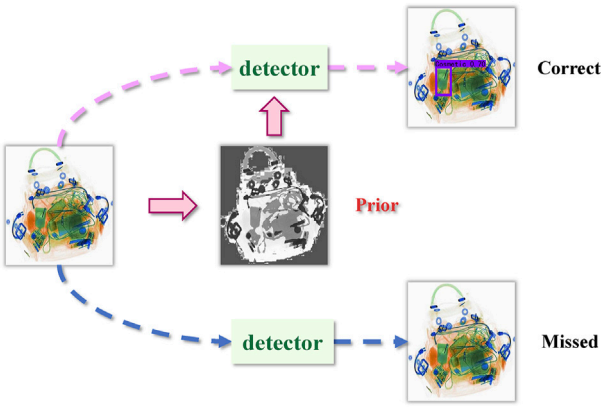


FIGURE 1
Various items in passengers' luggage and random permutations between articles result in cluttered X-ray images. For general object detectors, a large amount of irrelevant background information interference can easily lead to missed detections. With the assistance of the atomic number prior knowledge, our method can suppress background interference and detect items correctly.

Different from traditional detection tasks, in this scenario, there are various items in the passenger's luggage and random permutations between items, resulting in heavily cluttered X-ray images [1–4]. Therefore, object detection algorithms for general natural images do not perform well on cluttered X-ray images as in Figure 1. Fortunately, the tremendous success of deep learning [5–11] has made the intelligent detection of prohibited items possible by transforming it into an object detection task in computer vision [12–14]. Hence, many researchers have applied deep learning methods to prohibited object detection. Flitton et al. [15] explored 3D feature descriptors with application to threat detection in Computed Tomography (CT) airport baggage imagery. Bhowmik et al. [16] investigated the difference in detection performance achieved using real and synthetic X-ray training imagery for CNN architecture. Gaus et al; [17] evaluated several leading variants spanning the Faster R-CNN, Mask R-CNN, and RetinaNet architectures to explore the transferability of such models between varying X-ray scanners. Hassan et al; [18] presented a cascaded structure tensor framework that automatically extracts and recognizes suspicious items in multi-vendor X-ray scans. Zhao et al; [19] established the associations between feature channels and different labels and adjust the features according to the assigned labels (or pseudo labels) to tackle the overlapping object problem. These methods all improve detection performance to a certain extent but do not use the unique imaging characteristics of X-ray images to improve the algorithm.

Recently, some works have tried adding prior information about X-ray images to guide network learning, as shown in Figure 2 [20]. Obtained edge images by using the traditional edge detection algorithm Sobel. Chang et al. [4] found that different classes of prohibited objects have a clear distinction in physical size and used Otsu's threshold segmentation algorithm [21] to segment the original image into foreground and background, treating the foreground region as the approximate size of the detected object. Although these two methods improve the detection accuracy to a certain extent by introducing such prior information, the obtained prior information is easily disturbed by other irrelevant information due

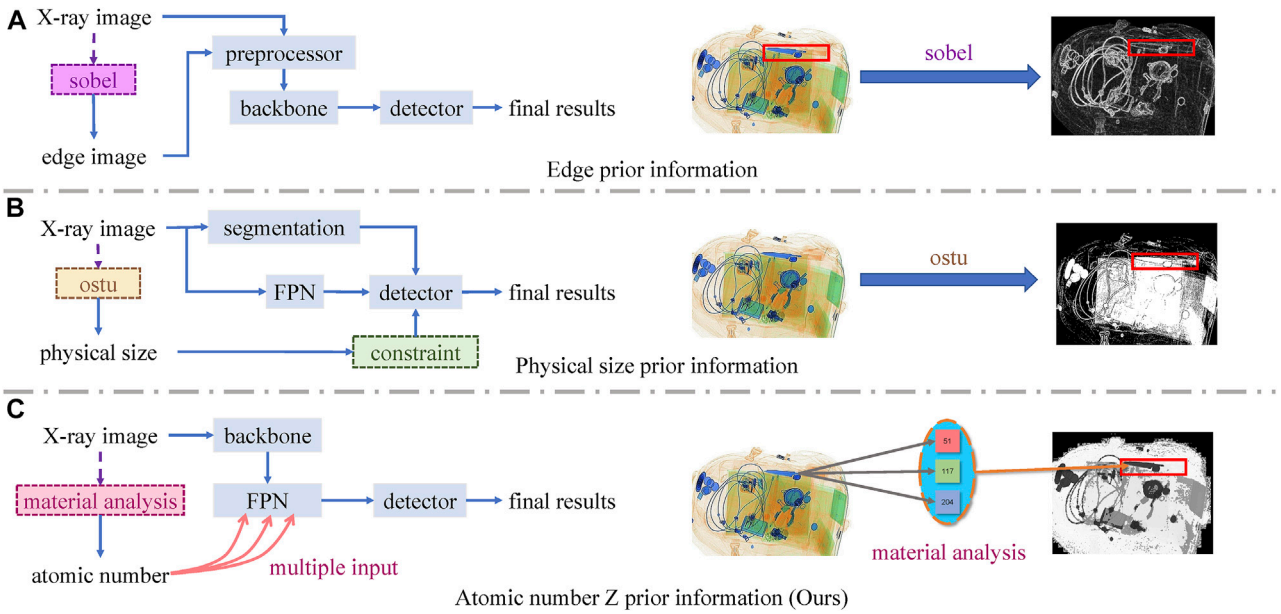


FIGURE 2
Framework comparisons between existing methods based on prior knowledge and our method. For each row, the left is the network framework, and the right is the visualization of prior knowledge. The prohibited objects in each X-ray image are annotated in red bounding boxes. **(A)** The method to obtain the boundary information of prohibited items will be seriously interfered with by the boundary information of unrelated items. **(B)** The way cannot fully believe the accuracy of treating the binarized foreground as the area of the detected object, especially when other items appear inside the detection box. **(C)** Unlike them, our method pays more attention to the atomic number feature, taking advantage of the distinction in atomic numbers to reduce the interference of useless background information.

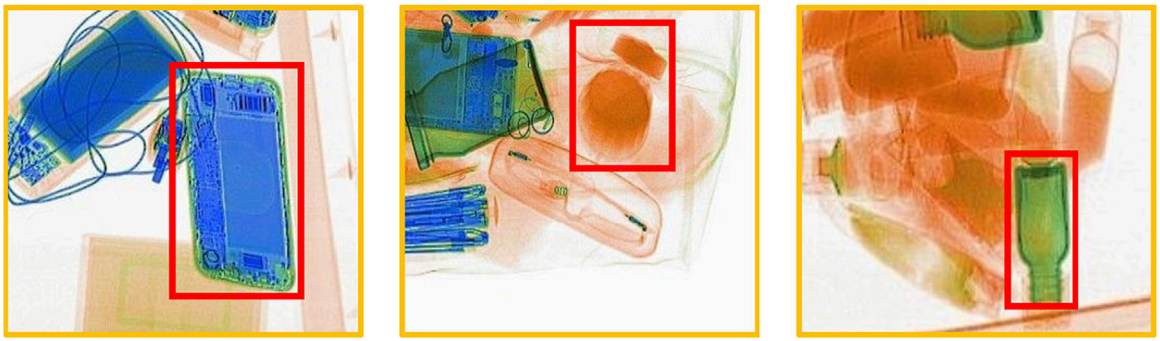


FIGURE 3
From left to right are inorganic matter, organic matter, and mixture.

to the messy distribution of prohibited items, which hinders further performance improvement. Specifically, in the presence of cluttered items, the former method to obtain the boundary information of prohibited terms is severely interfered with by the boundary information of irrelevant items. Furthermore, the latter cannot fully believe the accuracy of treating the binarized foreground as the area of the detected items, especially when other items appear inside the detection region.

In this paper, we propose a novel atomic number Z Prior Guided Network (ZPGNet) for heavily cluttered X-ray images, which can remove irrelevant background information by effectively incorporating the atomic number feature. Unlike optical images, X-ray images are generated by illuminating objects with X-ray. X-ray security inspection machine is based on the object difference in absorbing X-ray to detect the effective atomic number and then show distinct colors [22]. Specifically, the color information in X-ray images represents material information, where blue represents inorganic material, orange represents organic material, and green represents mixture [23], as shown in Figure 3. Atomic number images of X-ray image variants can directly reflect the material type of an item, which is the dominant information in X-ray images. This characteristic motivates us to explore this critical information to improve detection accuracy by removing irrelevant background information. Bhowmik et al; [24] examined the impact of atomic number images *via* the use of CNN architectures for the object detection task posed within X-ray baggage security screening and obviously illustrated a vital insight into the benefits of using atomic number images for object detection and segmentation tasks. However, they only simply connect atomic number images with RGB images and do not fully use atomic number images. In order to make full use of the atomic number features of items, we designed a Material Activation (MA) module. It cross-scale flows atomic number information through the network to mine deep material clues, which is beneficial to reduce irrelevant information interference in detecting prohibited items.

Atomic number images need to be collected manually, which increases the costs. In particularly, X-ray imaging systems render different materials in different colors. Blue represents inorganic material, orange represents organic material, and green represents mixture, as shown in Figure 3. Therefore, we can obtain the material classification of each pixel by analyzing the color. Thus, we propose an atomic number Z Prior Generation (ZPG) module, which automatically generates the atomic number feature according to the imaging color of X-ray images, as those shown in Figure 4.

Overall, the contributions of our work can be summarized as follows:

- We propose a novel atomic number Z Prior Guided Network (ZPGNet) to improve the detection accuracy of cluttered items by effectively incorporating the atomic number feature. In addition, the proposed method is generic and can be easily embedded into existing detection frameworks as a module.
- We propose an atomic number Z Prior Generation (ZPG) module, which automatically generates the atomic number feature according to the imaging color of X-ray images. Compared with the manual collection, the costs are significantly reduced.
- We design a Material Activation (MA) module to cross-scale fuse image features with the atomic number feature and then flow the fused features from high-level to low-level to enhance the ability of the model to mine deep material clues.
- We evaluate ZPGNet on the HiXray and OPIXray datasets and demonstrate that the performance of our ZPGNet is superior to state-of-the-art methods in identifying prohibited objects from cluttered X-ray baggage images.

2 Related work

In this section, we first introduce the existing public datasets for detecting prohibited items in X-ray images and then describe some generic object detection methods and some strategies to solve the clutter problem in X-ray images.

2.1 Security inspection image dataset

X-ray security inspection machines show different colors for different material items by the object distinction in absorption X-ray [22]. Therefore, it has many applications in many tasks, such as security inspection [4, 25–27]and medical imaging analysis [8, 28–33]. However, there are very few X-ray image datasets due to the particularity of security inspection scenes. To our knowledge, four recently published datasets are GDXray [22], SIXray [26], OPIXray

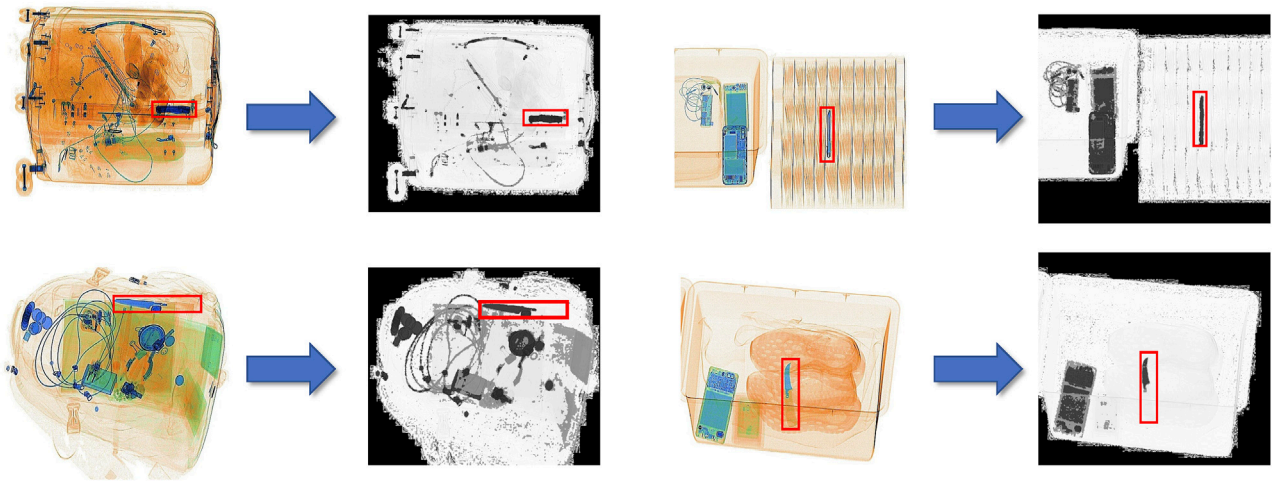


FIGURE 4
The X-ray image samples are from the OPIXray dataset. The left part of each set of photograph is the original image, and the right part is the atomic number image generated by our proposed ZPG method. The prohibited objects in each X-ray image are annotated in red bounding boxes.

[20], and HiXray [34]. The GDXray dataset has 19,407 images containing three prohibited items, namely, guns, darts, and razors. However, the GDXray dataset only contains grayscale images, which are far from realistic scenarios. The SIXray includes 1,059,231 X-ray images, which only have 8,929 labeled images. The pictures in the SIXray dataset are obtained by real security machines from several subway stations, which is more in line with the data distribution of real scenes. The OPIXray dataset is the first high-quality security target detection dataset, which contains five categories of prohibited items, namely, folding knives, straight knives, scissors, utility knives, and multitool knives, with a total of 8885 X-ray images. The HiXray dataset contains 44,364 X-ray images from daily security checks at international airports, which contain eight categories of prohibited items such as lithium batteries, liquids, and lighters that are common in daily life. Each image in the HiXray dataset is annotated by airport staff, which ensures the accuracy of the data.

2.2 Generic object detection

Object detection is an essential part of computer vision tasks, which supports many downstream tasks [35–38]. Methods based on convolutional neural networks can be summarized into two categories: single-stage [39–43] and multi-stage [44–46]. In recent years, compared with multi-stage detection methods, single-stage detection methods have been widely adopted due to their simple design and powerful performance. YOLOv3 [42] considers both real-time and accuracy by using the region proposal method. RetinaNet [41] improves the detection accuracy while maintaining the inference speed by solving the problem of class balance. It is far higher in real-time performance and accuracy than general multi-stage detection methods. FCOS [43] is anchor box free, as well as proposal free, to solve object detection in a per-pixel prediction fashion. In addition, YOLOv5 [47] makes several improvements based on YOLOv3, which significantly improves the detection speed and accuracy. However, so far, most object detection methods are for natural images. In the security check scene, various items in the passenger's luggage and random permutations between the objects resulted in heavily cluttered X-ray images, so the detection effect is often unperformed.

2.3 Solutions to heavily cluttered problems

Previous works have mainly focused on solving the problem of highly cluttered X-ray images. Shao et al. [48] proposed a foreground and background separation X-ray prohibited item detection framework that separates prohibited items from other items to exclude irrelevant background information. Tao et al. [34] proposed a lateral inhibition module to eliminate the influence of noisy neighboring regions on the interest object regions and activate the boundary of items by intensifying it.

3 Proposed method

Atomic number images of X-ray image variants can directly reflect item material, which is the dominant information in X-ray images. Inspired by this, we propose a novel atomic number Z Prior Guided Network (ZPGNet) for cluttered X-ray images, as shown in Figure 5. The ZPGNet consists of three main components: 1) an atomic number Z Prior Generation (ZPG) module automatically generates atomic number images, which reduces the cost of manually collecting atomic number images, 2) a Material Activation (MA) module fuses the atomic number feature to remove irrelevant background information, 3) a Bidirectional Enhancement (BE) module enriches feature expression through bidirectional information flow.

Specifically, we first design the ZPG module, combining the characteristics that different materials will show different colors, to map a three-channel (RGB) color image to a single-channel atomic number image. Then, we repeatedly pass the atomic number feature generated by the ZPG module into the network to pay more attention to item material information. To effectively fuse the extracted image features and the atomic number feature, MA cross-scale flows the atomic number feature under the extracted multi-scale features and uses a channel

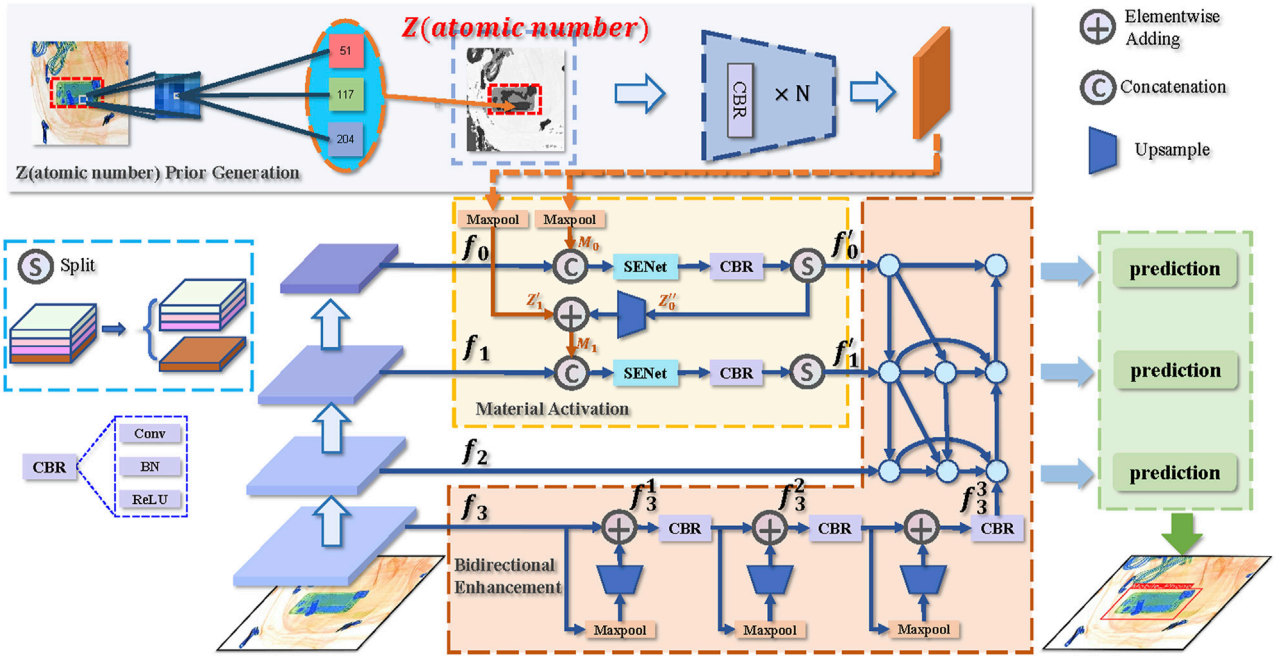


FIGURE 5
Overall framework of the proposed atomic number Z Prior Guided Network (ZPGNet). The network consists of three key modules, i.e., an atomic number Z Prior Generation (ZPG) module generating the atomic number feature, a Material Activation (MA) module cross-scale fusing the image features with the atomic number feature, and a Bidirectional Enhancement (BE) module mining contextual semantics for enhancing feature representation. CBR is composed of a convolution layer, a batch normalization layer, and a relu activation function. SENet stands for Squeeze-and-Excitation Networks [49].

attention module to self-adapt the importance of different features. Finally, we add a layer of low sampling rate features to obtain more detailed information and mine contextual semantics for enriching feature expression.

3.1 Z Prior Generation

Unlike optical images, X-ray images are generated by illuminating objects with X-rays, whose penetration is related to the material's density, size, and composition [22]. X-ray security machines detect the atomic number of objects based on the difference in absorbing X-rays, which then display a distinct color. Bhowmik et al. [24] proved that the introduction of atomic number images is an effective method to improve detection performance *via* large experiments. Inspired by this, the designed ZPG module compresses three-channel X-ray images into a single-channel to generate atomic number images that can highlight material differences. Compared with manually collecting atomic number images, it significantly reduced costs.

For each pixel in the RGB image, the maximum of the three channels will render its corresponding color. We use its subscripts to classify different materials.

$$g_{ij} = argmax\left(x_{ijk}\right) \tag{1}$$

where x_{ijk} denotes the value of the k-channel at position (i, j) the input image. *argmax* $(\bullet)$ denotes the index corresponding to finding the maximum value of an element.

Materials of the same class tend to present different depths of color due to different thicknesses. We introduce two variables, base-value B, and width-value W. The former is used to distinguish different materials, and the latter reflects the difference between the same materials.

$$B_{ij} = g_{ij} + \alpha \tag{2}$$

$$W_{ij} = \left(\sum x_{ij} - x_{ijg_{ij}}\right) * (1-\beta) * (1-\alpha) / (255+255) + x_{ijg_{ij}} * \beta * (1-\alpha) / 255 \tag{3}$$

Where α and β are hyperparameters that respectively control basis-value B and width-value W.

Finally, the basis-value B and width-value W are added and normalized, and then passed through a series of convolutional layers to obtain the atomic number feature Z.

$$Z_{ij} = \begin{cases} 0, & \text{if } x_{ij} = (255, 255, 255) \\ \left(B_{ij} + W_{ij}\right)/3, & \text{if others} \end{cases} \tag{4}$$

$$Z = \phi_n(Z) \tag{5}$$

where $\phi_n(\bullet)$ denotes the n-layer "Conv-BN-ReLU" operation, Since no items are in the white area, we specially treat for the pixel (255, 255, 255).

3.2 Material activation

In particular, different material items in X-ray images have clear distinctions according to their atomic number information, which is vital to suppress the interference of background information by mining deep material cues.

In cluttered X-ray images, the boundary and color information of prohibited items are easily interfered with by background

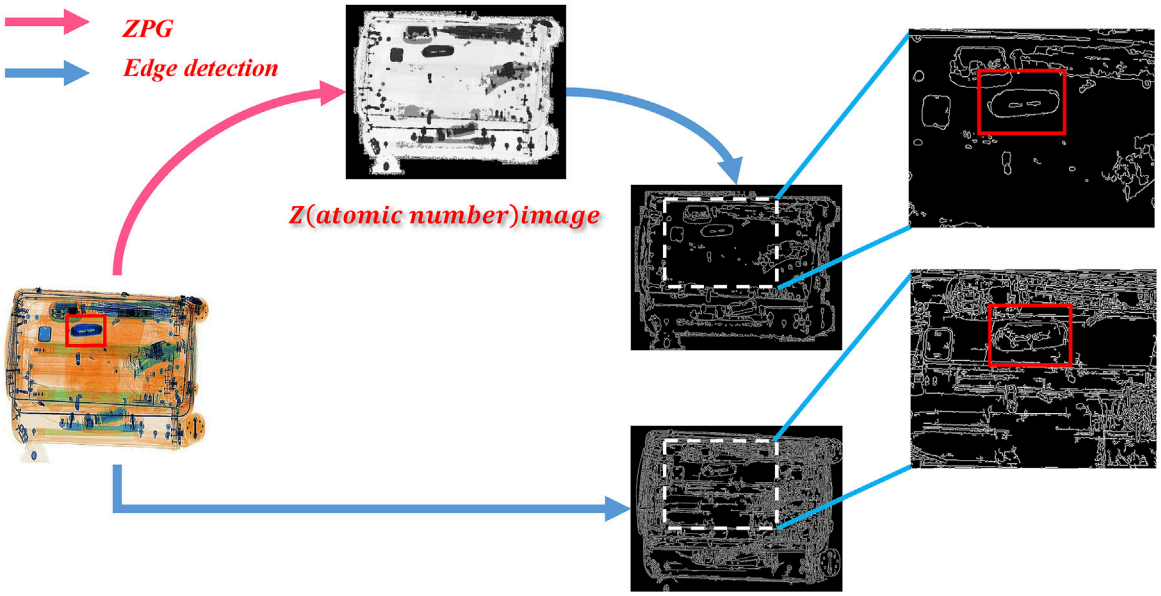


FIGURE 6
The bottom part shows the edge detection results obtained directly by the Canny algorithm [50], and the top part is obtained by first passing through the ZPG module and then through the Canny detection. It is intuitive to see that the edges of the items processed by the ZPG module are more evident than the original. The prohibited objects in each X-ray image are annotated in red bounding boxes.

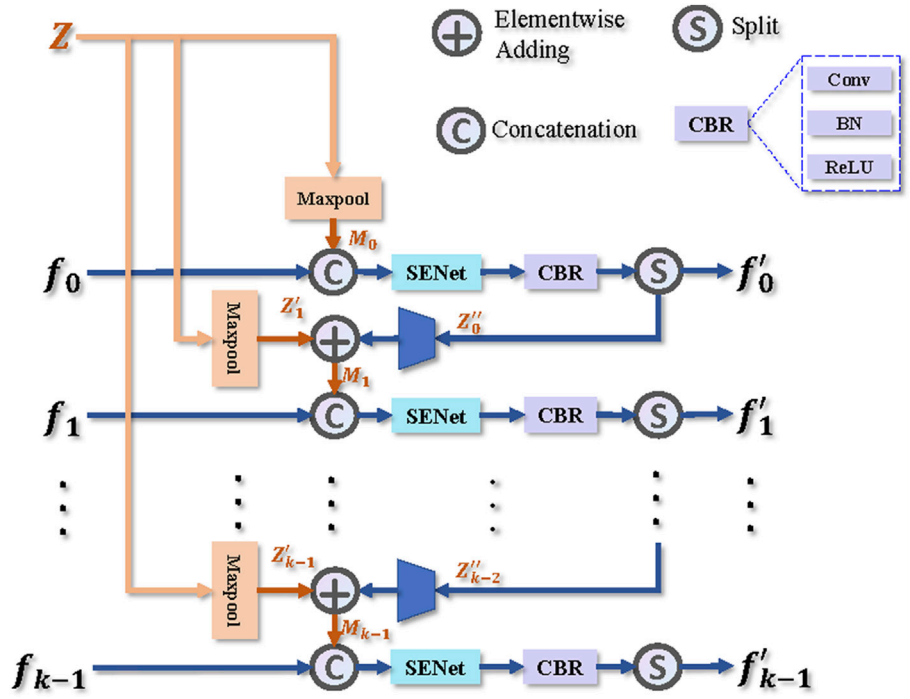


FIGURE 7
Illustration of the proposed Material Activation (MA) module, where k indicates that the input of the MA module has k different-scale feature maps.

information. MA introduces the atomic number feature to mine material cues, which is beneficial to reduce useless background information interference in detecting prohibited items, as shown in Figure 6.

Specifically, the backbone network has n feature map outputs $F = \{f_0, \ldots, f_{n-1}\}$. As shown in Figure 7, the MA structure makes the former k layers of F as the input. For Z and F feature maps, which are output by ZPG and Backbone, we pool the atomic number feature Z to increase the receptive field

and then add Z flowing down from the previous layer to get a more robust feature M. Furthermore, we concatenate them with F for information fusion and apply channel attention operation (Squeeze-and-Excitation Networks [49]) $SE(\bullet)$ on the fused features to adapt the importance between the material feature and other image features (edge, texture, size, etc.).

$$Z_i' = \mathcal{D}_i(Z) \tag{6}$$

$$F_{ei} = \phi_1\left(SE\left(f_i \| M_i\right)\right) \tag{7}$$

where $\|$ represents the operation of concatenating, $\mathcal{D}_i(\bullet)$ denotes the Pooling operation.

Separate F_{ei} into f_i' and $Z_i^{''}$ along the channel dimension, whose dimensions are the same as f_i and Z_i', respectively, where the f_i' is used as the input of the next BE module, and the $Z_i^{''}$ is passed to the next layer of the MA module as an enhanced atomic number feature to obtain the more robust feature.

$$\begin{cases} f_i' = F_{ei}^0 \\ Z_i^{''} = F_{ei}^1 \end{cases} \tag{8}$$

$$M_i = Z_i' + \mathcal{U}\left(Z_{i-1}^{''}\right) \tag{9}$$

where F_{ei}^0 and F_{ei}^1 denote the two features obtained by separating F_{ei} along the channel, $\mathcal{U}(\bullet)$ denotes the Upsample operation. Especially, $M_0 = Z_0'$.

3.3 Bidirectional Enhancement

When the down-sampling rate is high, it is easy to obtain larger receptive fields and more large-scale item information, which is beneficial for detecting large-scale prohibited objects. However, for some minor prohibited items, too large a downsampling rate tends to lose too much detail feature information of small-scale objects.

In the HiXray [34] high-quality prohibited items dataset, the average resolution of images is 1,200*900, with the largest resolution being 2000*1,024. The resolution of some small lighters is only 21*57, which is about 1/1,000 the size of the original image. After excessive downsampling, the feature information of lighters is seriously missing, resulting in poor detection in SSD [51], LIM [34], DOAM [20], and other detection models.

BE module adds a low sampling rate feature to obtain more detailed information about the tiny-size prohibited items. However, the low sampling rate feature often contains additional noise information. We remove noisy information by performing multiple pooling operations.

$$f_3^{i+1} = \phi_1\left(\mathcal{U}\left(\mathcal{D}_i\left(f_3^i\right)\right) + f_3^i\right) \tag{10}$$

where f_3^3 is the finally denoised low-sampling rate feature, and specifical $f_3^0 = f_3$.

Finally, the material activation feature $\{f_0', \ldots, f_{k-1}'\}$ obtained by the MA module, Backbone output feature $\{f_k, \ldots, f_2\}$, and f_3^3 are streamed bidirectionally, which mines contextual semantics to enrich feature expression.

4 Experiments

4.1 Datasets and evaluation Metrics

We conduct extensive experiments to evaluate our proposed model on two prohibited item detection datasets, HiXray [34] and OPIXray [20]. HiXray dataset consists of 45,364 X-ray images from routine security checks at international airports, which contains 8 categories of 102,928 everyday prohibited items commonly seen in daily life, such as lithium batteries, liquids, lighters, etc. Each image in the HiXray dataset was annotated by an airport employee, which ensures the accuracy of the data. OPIXray dataset is the first high-quality object detection dataset for security, which focused on the widely-occurred prohibited item "cutter", annotated manually by professional inspectors from the international airport. The dataset contains five categories of prohibited objects with a total of 8885 X-ray images (7,109 for training and 1,776 for testing).

Average Precision (AP) denotes the area under the precision-recall curve of the detection results for a single category of objects. To fairly evaluate the performance of all models, we compute the mean average precision (mAP) with an IOU threshold of .5. In addition, we calculate AP for all categories for each model to see the improvement for each category.

4.2 Implementation details

All our experiments were done in Pytorch and trained on one NVIDIA RTX 3090 GPU with the initial learning rate set to 1e-2. The parameters were optimized through stochastic gradient descent (SGD). The momentum and weight decay are set to .937 and .0005, respectively. Besides, two new hyperparameters were introduced with respect to the module ZPG, i.e., α and β, which respectively control base-value B and width-value W, and values are set to .4 and .5.

4.3 Quantitative results

We test the model performance on HiXray [34] and OPIXray [20] datasets. Specifically, we embedded ZPGNet into YOLOv3 [42] and YOLOv5s [47] and compared it with the state-of-the-art methods DOAM [20] and LIM [34]. Table 1 presents the experimental results of DOAM, LIM, and the proposed ZPGNet on HiXray and OPIXray datasets. In order to illustrate the effectiveness of our method and better compare it with the existing state-of-the-art (SOTA) models, we use YOLOv3 and YOLOv5s as this baseline.

4.3.1 Results on HiXray dataset

The experimental results of different algorithms on the HiXray [34] dataset are shown in Table 1. For a fair comparison, we adopt the same baseline YOLOv5s [47] as DOAM [20] and LIM [34], which performs the best results on both DOMA and LIM. The proposed method ZPGNet with YOLOv5s baseline improves to 83.9% in mean average prediction, outperforming DOAM and LIM by 1.7% mAP_{50} and .7% mAP_{50}, respectively. In order to further verify the effectiveness of our model, we also adopted the YOLOv3 [42] baseline, which is still 1.2% mAP_{50} higher than the SOTA method (YOLOv5s + LIM).

The (YOLOv3+ZPGNet) experiment results show that our method is lower than some methods in some categories Water, Laptop, Mobile Phone, and Tablet, but has an 8.0% AP and 4.8% AP improvement in the cosmetics and lighter categories, respectively, compared to the SOTA method LIM. Cosmetics belong to the mixtures category, commonly disturbed by organic substances (such as plastics), resulting in decreased detection confidence or even missed detection. The significant

TABLE 1 Quantitative evaluation results on the HiXray dataset and OPIXray dataset. Where PO1, PO2, WA, LA, MP, TA, CO, and NL denote "Portable Charger 1 (lithium-ion prismatic cell)", "Portable Charger 2 (lithium-ion cylindrical cell)", "Water," "Laptop," "Mobile Phone," "Tablet," "Cosmetic" and "Non-metallic Lighter" in the HiXray dataset. FO, ST, SC, UT, and MU donate "Folding Knife," "Straight Knife," "Scissor," "Utility Knife," and "Multi-tool Knife" in the OPIXray dataset, respectively.

Method	HiXray									OPIXray					
	mAP_{50}	PO1	PO2	WA	LA	MP	TA	CO	NL	mAP_{50}	FO	ST	SC	UT	MU
SSD [51]	71.4	87.3	81.0	83.0	97.6	93.5	92.2	36.1	.01	70.9	76.9	35.0	93.4	65.9	83.3
SSD + DOAM [20]	72.1	88.6	82.9	83.6	97.5	94.1	92.1	38.2	.01	74.0	81.4	41.5	95.1	68.2	83.8
SSD + LIM [34]	73.1	89.1	84.3	84.0	97.7	92.4	92.4	42.3	0.1	74.6	81.4	42.4	95.9	71.2	82.1
Xdet [4]	—	—	—	—	—	—	—	—	—	86.7	90.4	76	91.5	84.3	91.3
FCOS [43]	75.7	88.6	86.4	86.8	89.9	88.9	88.9	63.0	13.3	82.0	86.4	68.5	90.2	78.4	86.6
FCOS + DOAM [20]	76.2	88.6	87.5	87.8	89.9	89.7	88.8	63.5	12.7	82.4	86.5	68.6	90.2	78.8	87.7
FCOS + LIM [34]	77.3	88.9	88.2	88.3	90.0	89.8	89.2	69.8	14.4	83.1	86.6	71.9	90.3	79.9	86.8
ATSS [19]	—	—	—	—	—	—	—	—	—	86.6	92.3	72.0	96.6	80.38	91.7
ATSS + DOAM [19]	—	—	—	—	—	—	—	—	—	85.6	90.7	66.8	96.2	81.8	92.5
ATSS + Lacls [19]	—	—	—	—	—	—	—	—	—	88.3	90.0	75.0	97.6	85.7	93.0
YOLOv5s [47]	81.7	95.5	94.5	92.8	97.9	98.0	94.9	63.7	16.3	87.8	93.4	67.9	98.1	85.4	94.1
YOLOv5s + DOAM [20]	82.2	95.9	94.7	93.7	98.1	98.1	95.8	65.0	16.1	88.0	93.3	69.3	97.9	84.4	95.0
YOLOv5s + LIM [34]	83.2	96.1	95.1	**93.9**	**98.2**	**98.3**	**96.4**	65.8	21.3	90.6	94.8	77.6	**98.2**	**88.9**	93.8
YOLOv5s + ZPGNet (Ours)	83.9	95.7	**95.2**	92.5	96.5	97.7	94.4	66.4	**33.0**	**90.7**	**95.0**	**79.3**	98.0	86.8	94.2
YOLOv3 [42]	83.0	**96.7**	94.9	91.9	97.9	97.7	94.0	71.9	18.6	78.2	92.5	36.0	97.3	70.8	**94.4**
YOLOv3+ZPGNet (Ours)	**84.4**	96.6	95.2	92.7	97.7	98.0	95.2	**73.8**	26.1	85.4	88.5	65.1	96.7	83.5	93.3

Bold values represent the best performance in the same evaluation index.

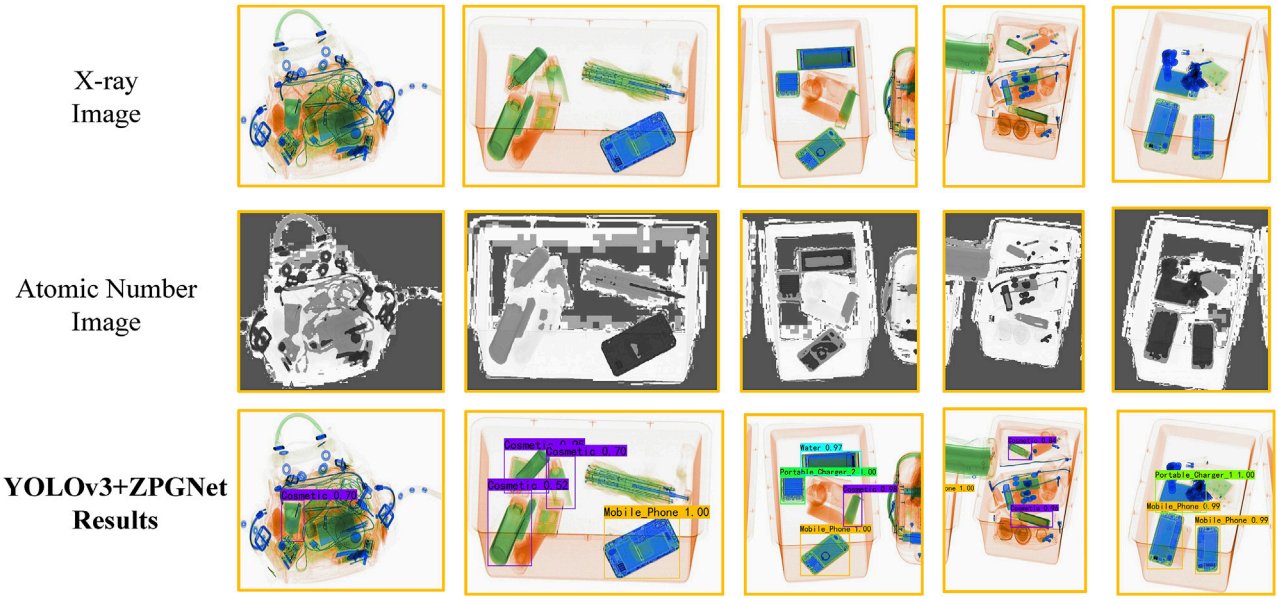


FIGURE 8
Visualizations of the original images, atomic number images, and detection results of the ZPGNet-integrated model. Our proposed ZPGNet uses atomic images to pay more attention to material information and thus achieve better performance.

improvement in cosmetics indicates that our method, introducing the atomic number feature map, can better reduce the interference of useless information in Figure 8. This advantage is facilitated by our method of paying extra attention to the material information using atomic number features. Lighters in luggage are tiny in size and prone to profound feature loss after downsampling. Our

TABLE 2 Comparisons between the ZPGNet-integrated network and three object detection methods.

Method	mAP_{50}	FO	ST	SC	UT	MU
RetinaNet [41]	87.4	89.4	69.2	98.2	**86.3**	**94.0**
RetinaNet + ZPGNet	**88.1**	**91.3**	**72.1**	**98.7**	85.8	92.6
YOLOv5s [47]	87.8	93.4	67.9	**98.1**	85.4	94.1
YOLOv5s + ZPGNet	**90.7**	**95.0**	**79.3**	98.0	**86.8**	**94.2**
YOLOv3 [42]	78.2	**92.5**	36.0	**97.3**	70.8	**94.4**
YOLOv3+ZPGNet	**85.4**	88.5	**65.1**	96.7	**83.5**	93.3

We embedded our method into three different baseline models respectively and divided the models embedded with and without our method into a group, where the bold figures represent the best performance in a group.

method achieves 11.7% *AP* improvement over LIM [34] with the same baseline YOLOv5s in the lighter category, which is due to the fact that we use a low sampling rate feature map in the BE module to increase the information of small prohibited items.

4.3.2 Results on OPIXray dataset

Table 1 represents the performance of our method on the OPIXray [20] dataset. With the same baseline YOLOv5s [47], ZPGNet outperforms DOAM [20] and LIM [34] by 2.7% mAP_{50} and .1% mAP_{50}, respectively. In particular, ZPGNet has the highest score on mAP_{50} among all the models. It can be clearly seen that the proposed method ZPGNet achieves significant performance improvement based on YOLOv3 [42], especially on *AP* of the severely occluded prohibited items named "straight knife" improved by 29.1%. This benefits from the fact that our method effectively removes the interference of irrelevant background information.

4.4 Generality verification

To further evaluate the effectiveness of the proposed model ZPGNet and verify that ZPGNet can be applied to various detection networks, we choose the classical detection models YOLOv3 [42], RetinaNet [41], and YOLOv5s [47] to use our method. Experiments were performed on the OPIXray dataset [20]. As shown in Table 2, our approach ZPGNet improves YOLOv3 by 7.2% mAP_{50}, RetinaNet by .7% mAP_{50}, and YOLOv5s by 2.9% mAP_{50}, respectively. Many objects are commonly disturbed by useless items, quickly resulting in low confidence or even miss detection on the general detection model. As shown in Figure 9, the comparison plot of the experimental results in the first and second rows shows that even with high confidence, there is a particular improvement after introducing the atomic number features. Embedding ZPGNet makes the network pay more attention to object material information to reduce the interference of ineffective information and alleviate the problems of low confidence and missed detection. This indicates that our model can be embedded into most detection networks as a plug-and-play component to minimize the interference of useless background information and achieve better performance.

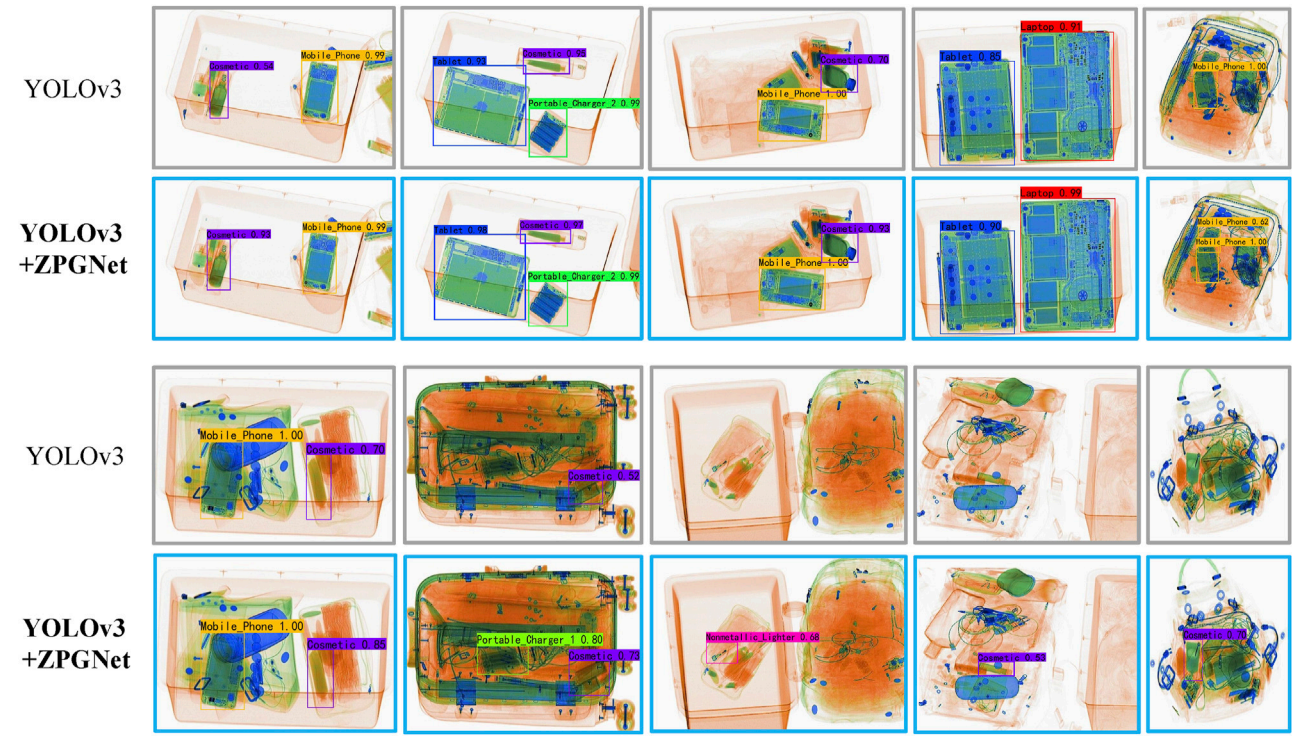


FIGURE 9
Visual results of both the baseline YOLOv3 and the ZPGNet-integrated model. There are many missed and low-confidence prohibited items in baseline YOLOv3. After embedding the proposed ZPGNet, the ability to detect items has been significantly improved, especially for heavily cluttered X-ray images.

TABLE 3 Ablation results of the proposed ZPG, MA, and BE on the HiXray dataset.

Method	mAP_{50}	PO1	PO2	WA	LA	MP	TA	CO	NL
YOLOV5s [47]	81.7	95.5	94.5	**92.8**	**97.9**	**98.0**	**94.9**	63.7	16.3
+ZPG + MA	83.1	95.3	**95.5**	92.4	94.9	97.7	93.6	**69.0**	26.0
+ZPG + MA + BE	**83.9**	**95.7**	95.2	92.5	96.5	97.7	94.4	66.4	**33.0**

Bold values represent the best performance in the same evaluation index.

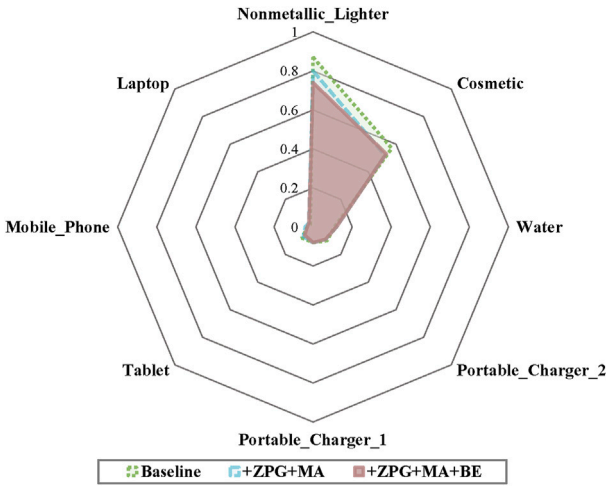


FIGURE 10
Performance comparison of different categories. The number on the gray line indicates the log-average miss rate. Useless background information interference can easily lead to prohibited item missed detections. With the proposed ZPG, MA, and BE, the log-average miss rate of prohibited items (i.e., cosmetic and lighter) is significantly reduced.

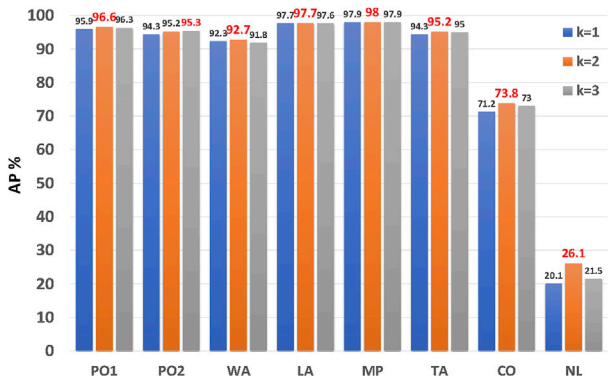


FIGURE 11
Bar graph of AP variation of all categories corresponding to different layers number MA module.

4.5 Ablation study

In this subsection, we conduct a series of ablation experiments to analyze the influence of involved hyperparameters and the contribution of critical components of the proposed ZPGNet. In the ablation study, all experiments were performed on the HiXray dataset [34].

4.5.1 Effectiveness of ZPG, MA, and BE

ZPG, MA, and BE are essential modules in ZPGNet, and we embed them one by one into YOLOv5s [47] to evaluate their performance. The insertion of ZPG requires the support of MA, so unity emplaces ZPG and MA together into the model. All experiments here uniformly set the number of MA layers to 2. As shown in Table 3, the network embedded with ZPG and MA modules improves its performance by 1.4% mAP_{50} compared to the base model, especially in the cosmetics category, where it improves by 5.3% mAP_{50}. Cosmetics are commonly disturbed by organic substances (such as plastics), resulting in low confidence and missed detection. The significant improvement in cosmetics indicates that our method, introducing the atomic number features, can better reduce the interference of useless information, as shown in Figure 10. After applying the Bidirectional Enhancement (BE) module, the performance is 2.2% mAP_{50} higher than the basic module and .8% mAP_{50} higher than that embedded with MA and ZPG, which proves the effectiveness of the BE module.

4.5.2 Number of layers in MAs

We also show the effects of different layer numbers in the proposed MA, as shown in Figure 11. The model performs best when the layer numbers equal 2. The excessive number of layers can lead to performance degradation of the MA module. We believe that the possible reason is that the over-introduction of the atomic number feature leads to the suppression of other essential cues, which leads to a degradation in performance. When MA layers are equal to 2, it can well balance the importance between the atomic number feature and other features. So, in other experiments, we set the layer numbers in each MA to 2.

5 Conclusion

Prohibited item detection in X-ray images is an effective measure to maintain public safety. The interference of a large amount of useless background information caused by object disordered placement is an urgent problem to be addressed in prohibited item detection. Inspired by the imaging characteristics of X-ray images, this paper proposes an atomic number Z Prior Generation (ZPG) method, which can automatically generate atomic number images and reduce the cost of manual acquisition. Furthermore, we designed an atomic number Z Prior Guided Network (ZPGNet) to solve useless background information interference in prohibited item detection. The

proposed ZPGNet method cross-scale flows the atomic number Z information through the network to mine deep material clues to reduce irrelevant background information interference. We comprehensively evaluate ZPGNet on HiXray and OPIXray datasets, and this result shows that ZPGNet can be embedded into most detection networks as a plug-and-play module and achieve higher performance. There is still a severe occlusion problem in X-ray images, but this paper does not solve the occlusion problem. In the future, we intend to use features such as contour and scale to solve the occlusion problem between items.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/OPIXray-author/OPIXray.

Author contributions

Conceptualization, JC, JL, MM, XG, and SG; methodology, JC; software, MM, and SG; validation, JC; investigation, JL and SG; writing—original draft preparation, JC and JL; writing—review and editing, XG, MM, and SG; visualization, JC; funding acquisition, JL and XG. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grants No. 62102057 and No. 62036007, in part by the Natural Science Foundation of Chongqing under Grand No. CSTB2022NSCQ-MSX1024, in part by the Chongqing Postdoctoral Innovative Talent Plan under Grant No. CQBX202217, in part by the Postdoctoral Science Foundation of China under Grant No. 2022M720548, in part by the Special Project on Technological Innovation and Application Development under Grant No. cstc2020jscx-dxwtB0032, and in part by Chongqing Excellent Scientist Project under Grant No. cstc2021ycjh-bgzxm0339.

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. Gaus YFA, Bhowmik N, Akçay S, Guillén-Garcia PM, Barker JW, Breckon TP. Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered x-ray security imagery. In: 2019 international joint conference on neural networks (IJCNN); July 14-19, 2019; Budapest, Hungary (2019).

2. Hassan T, Bettayeb M, Akçay S, Khan S, Bennamoun M, Werghi N. Detecting prohibited items in x-ray images: A contour proposal learning approach. In: 2020 IEEE International Conference on Image Processing (ICIP); October 25-28, 2020 (2020).

3. Isaac-Medina BK, Willcocks CG, Breckon TP. Multi-view object detection using epipolar constraints within cluttered x-ray security imagery. In: 2020 25th International Conference on Pattern Recognition (ICPR); 10 - 15 January 2021; ITALY (2021). p. 9889.

4. Chang A, Zhang Y, Zhang S, Zhong L, Zhang L. Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. *Knowledge-Based Syst* (2022) 237:107916. doi:10.1016/j.knosys.2021.107916

5. Xia S, Wang F, Xie F, Huang L, Wang Q, Ling X. An efficient and robust target detection algorithm for identifying minor defects of printed circuit board based on phfe and fl-rfcn. *Front Phys* (2021) 9:661091. doi:10.3389/fphy.2021.661091

6. Franzel T, Schmidt U, Roth S. Object detection in multi-view x-ray images. In: *Joint DAGM (German association for pattern recognition) and OAGM symposium*. Charm: Springer (2012). p. 144–54.

7. Gao R, Sun Z, Huyan J, Li W, Xiao L, Yao B, et al. Small foreign metal objects detection in x-ray images of clothing products using faster r-cnn and feature pyramid network. *IEEE Trans Instrumentation Meas* (2021) 70:1–11. doi:10.1109/tim.2021.3077666

8. Luz E, Silva P, Silva R, Silva L, Guimarães J, Miozzo G, et al. Towards an effective and efficient deep learning model for Covid-19 patterns detection in x-ray images. *Res Biomed Eng* (2022) 38:149–62. doi:10.1007/s42600-021-00151-6

9. Akcay S, Breckon T. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition* (2022) 122:108245. doi:10.1016/j.patcog.2021.108245

10. Tian C, Fei L, Zheng W, Xu Y, Zuo W, Lin C-W. Deep learning on image denoising: An overview. *Neural Networks* (2020) 131:251–75. doi:10.1016/j.neunet.2020.07.025

11. Goudet O, Grelier C, Hao J-K. A deep learning guided memetic framework for graph coloring problems. *Knowledge-Based Syst* (2022) 258:109986. doi:10.1016/j.knosys.2022.109986

12. Wei X, Liu S, Xiang Y, Duan Z, Zhao C, Lu Y. Incremental learning based multi-domain adaptation for object detection. *Knowledge-Based Syst* (2020) 210:106420. doi:10.1016/j.knosys.2020.106420

13. Pérez-Hernández F, Tabik S, Lamas A, Olmos R, Fujita H, Herrera F. Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. *Knowledge-Based Syst* (2020) 194:105590. doi:10.1016/j.knosys.2020.105590

14. Zuo S, Xiao Y, Chang X, Wang X. Vision transformers for dense prediction: A survey. *Knowledge-Based Syst* (2022) 253:109552. doi:10.1016/j.knosys.2022.109552

15. Flitton G, Breckon TP, Megherbi N. A comparison of 3d interest point descriptors with application to airport baggage object detection in complex ct imagery. *Pattern Recognition* (2013) 46:2420–36. doi:10.1016/j.patcog.2013.02.008

16. Bhowmik N, Wang Q, Gaus YFA, Szarek M, Breckon TP (2019). The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited x-ray imagery. arXiv preprint arXiv:1909.11508

17. Gaus YFA, Bhowmik N, Akcay S, Breckon T. Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within x-ray security imagery. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA); December 16-19, 2019; Boca Raton, Florida, USA (2019). p. 420–5.

18. Hassan T, Khan SH, Akcay S, Bennamoun M, Werghi N (2019). Cascaded structure tensor framework for robust identification of heavily occluded baggage items from multi-vendor x-ray scans. *arXiv preprint arXiv:1912.04251*

19. Zhao C, Zhu L, Dou S, Deng W, Wang L. Detecting overlapped objects in x-ray security imagery by a label-aware mechanism. *IEEE Trans Inf Forensics Security* (2022) 17:998–1009. doi:10.1109/tifs.2022.3154287

20. Wei Y, Tao R, Wu Z, Ma Y, Zhang L, Liu X. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In: Proceedings of the 28th ACM International Conference on Multimedia; October 12 - 16, 2020; Seattle WA USA (2020). p. 138–46.

21. Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions Systems, Man, Cybernetics* (1979) 9:62–6. doi:10.1109/tsmc.1979.4310076

22. Mery D. *Computer vision for x-ray testing*, 10. Switzerland: Springer International Publishing (2015). p. 973–8.

23. Liang KJ, Heilmann G, Gregory C, Diallo SO, Carlson D, Spell GP, et al. Automatic threat recognition of prohibited items at aviation checkpoint with x-ray imaging: A deep learning approach. *Anomaly Detect Imaging X-Rays (Adix) (Spie)* (2018) 10632:1063203. doi:10.1117/12.2309484

24. Bhowmik N, Gaus YFA, Breckon TP. On the impact of using x-ray energy response imagery for object detection via convolutional neural networks. In: 2021 IEEE International Conference on Image Processing (ICIP); 19-22 September, 2021; Alaska, USA (2021). p. 1224.

25. Viriyasaranon T, Chae S-H, Choi J-H. Mfa-net: Object detection for complex x-ray cargo and baggage security imagery. *Plos one* (2022) 17:e0272961. doi:10.1371/journal.pone.0272961

26. Miao C, Xie L, Wan F, Su C, Liu H, Jiao J, et al. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 20 2021 to June 25 2021; Nashville, TN, USA. (2019). p. 2119–28.

27. Mery D, Riffo V, Zscherpel U, Mondragón G, Lillo I, Zuccar I, et al. Gdxray: The database of x-ray images for nondestructive testing. *J Nondestructive Eval* (2015) 34:42–12. doi:10.1007/s10921-015-0315-7

28. Talamonti C, Kanxheri K, Pallotta S, Servoli L. Diamond detectors for radiotherapy x-ray small beam dosimetry. *Front Phys* (2021) 9:632299. doi:10.3389/fphy.2021.632299

29. Fourcade A, Khonsari R. Deep learning in medical image analysis: A third eye for doctors. *J stomatology, Oral Maxill Surg* (2019) 120:279–88. doi:10.1016/j.jormas.2019.06.002

30. Zhou Z, Sodha V, Rahman Siddiquee MM, Feng R, Tajbakhsh N, Gotway MB, et al. Models Genesis: Generic autodidactic models for 3d medical image analysis. In: International conference on medical image computing and computer-assisted intervention; September 18th to 22nd 2022; Singapore (2019). p. 384–93.

31. Yang Y, Yan T, Jiang X, Xie R, Li C, Zhou T. Mh-net: Model-data-driven hybrid-fusion network for medical image segmentation. *Knowledge-Based Syst* (2022) 248:108795. doi:10.1016/j.knosys.2022.108795

32. Tang P, Yang P, Nie D, Wu X, Zhou J, Wang Y. Unified medical image segmentation by learning from uncertainty in an end-to-end manner. *Knowledge-Based Syst* (2022) 241:108215. doi:10.1016/j.knosys.2022.108215

33. Liu Y, Wang H, Chen Z, Huangliang K, Zhang H. TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation. *Knowledge-Based Syst* (2022) 256:109859. doi:10.1016/j.knosys.2022.109859

34. Tao R, Wei Y, Jiang X, Li H, Qin H, Wang J, et al. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 27 October - 2 November 2019; Seoul, South Korea (2021). p. 10923.

35. 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, Las Vegas, NV, June 27–30, 2016 (2016). 779–88.

36. Zhao Z-Q, Zheng P, Xu S-t., Wu X. Object detection with deep learning: A review. *IEEE Trans Neural networks Learn Syst* (2019) 30:3212–32. doi:10.1109/tnnls.2018.2876865

37. Zou Z, Shi Z, Guo Y, Ye J (2019). Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*

38. Velayudhan D, Hassan T, Damiani E, Werghi N. Recent advances in baggage threat detection: A comprehensive and systematic survey. *ACM Comput Surv* (2022). doi:10.1145/3549932

39. Zheng W, Tang W, Jiang L, Fu C-W. Se-ssd: Self-ensembling single-stage object detector from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 20 2021 to June 25 2021; Nashville, TN, USA (2021). p. 14494–503.

40. Zhang Y, Xie F, Huang L, Shi J, Yang J, Li Z. A lightweight one-stage defect detection network for small object based on dual attention mechanism and pafpn. *Front Phys* (2021) 9:708097. doi:10.3389/fphy.2021.708097

41. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision; October 13 - 16, 2003; NW Washington, DC, United States (2017). p. 2980–8.

42. Redmon J, Farhadi A (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*

43. Tian Z, Shen C, Chen H, He T. Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision; Oct. 27 2019 to Nov. 2 2019; Seoul, Korea (2019). p. 9627.

44. Ouyang W, Luo P, Zeng X, Qiu S, Tian Y, Li H, et al. (2014). Deepid-net: Multi-stage and deformable deep convolutional neural networks for object detection. arXiv preprint arXiv:1409.3505

45. Du L, Zhang R, Wang X. Overview of two-stage object detection algorithms. *J Phys Conf Ser* (2020) 1544:012033. doi:10.1088/1742-6596/1544/1/012033

46. Xie X, Cheng G, Wang J, Yao X, Han J. Oriented r-cnn for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 11-17 Oct. 2021 (2021). p. 3520–9.

47. Jocher G, Stoken A, Borovec J, Nano Code012 C, Changyu L, Wang M, et al. (2020). ultralytics/yolov5: v3

48. Shao F, Liu J, Wu P, Yang Z, Wu Z. Exploiting foreground and background separation for prohibited item detection in overlapping x-ray images. *Pattern Recognition* (2022) 122:108261. doi:10.1016/j.patcog.2021.108261

49. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 18-20 June 1996; San Francisco, CA, USA (2018). p. 7132.

50. Ding L, Goshtasby A. On the canny edge detector. *Pattern recognition* (2001) 34:721–5. doi:10.1016/s0031-3203(00)00023-6

51. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: Single shot multibox detector. In: European conference on computer vision; 8-16 October; Amsterdam, The Netherlands (2016). p. 21–37.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fphy.2022.1101923

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Jinxing Li,
Harbin Institute of Technology, Shenzhen, China
Guanqiu Qi,
Buffalo State College, United States

*CORRESPONDENCE
Yingying Su,
yy_su2000@163.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 18 November 2022
ACCEPTED 13 December 2022
PUBLISHED 05 January 2023

CITATION
He Y, Su Y, Wang X, Yu J and Luo Y (2023), An improved method MSS-YOLOv5 for object detection with balancing speed-accuracy.
Front. Phys. 10:1101923.
doi: 10.3389/fphy.2022.1101923

An improved method MSS-YOLOv5 for object detection with balancing speed-accuracy

Yaping He[1], Yingying Su[1]*, Xiaofeng Wang[2], Jun Yu[1] and Yu Luo[1]

[1]College of Electrical Engineering, Chongqing University of Science and Technology, Chongqing, China, [2]College of Mathematical and Physical Sciences, Chongqing University of Science and Technology, Chongqing, China

For deep learning-based object detection, we present a superior network named MSS-YOLOv5, which not only considers the reliability in complex scenes but also promotes its timeliness to better adapt to practical scenarios. First of all, multi-scale information is integrated into different feature dimensions to improve the distinction and robustness of features. The design of the detectors increases the variety of detection boxes to accommodate a wider range of detected objects. Secondly, the pooling method is upgraded to obtain more detailed information. At last, we add the Angle cost and assign new weights to different loss functions to accelerate the convergence and improve the accuracy of network detection. In our network, we explore four variants MSS-YOLOv5s, MSS-YOLOv5m, MSS-YOLOv5x, and MSS-YOLOv5l. Experimental results of MSS-Yolov5s show that our technique improves mAP on the PASCAL VOC2007 and PASCAL 2012 datasets by 2.4% and 2.9%, respectively. Meanwhile, it maintains a fast inference speed. At the same time, the other three models have different degrees of performance improvement in terms of balancing speed and precision in challenging detection regions.

KEYWORDS

multi-scale fusion, YOLOv5, loss function, softpool, object detection

1 Introduction

With the rapid development of science and technology, object detection technology has become a hot research problem [1]. Object detection has been useful in enhancing production and life efficiency in a variety of industries, including intelligent transportation, steel defect identification, face detection, and others. In terms of smart transportation, A great many traffic accidents happened in the world because of fatigued driving and drunk driving. Globally more than 1.25 million people died in traffic accidents and economic losses amount to billions of dollars every year. Due to the increasing number of vehicles and the irregular operation of drivers, the accident rate is further increasing, which brings many adverse effects to our production life. The computer-aided driving system monitors and senses the surrounding environment through deep learning

algorithms, and transmits information about obstacles in front of the vehicle to the driver or driverless system to facilitate the next effective operation, which is of great importance to reducing the incidence of traffic accidents. For steel defect detection, numerous steel varieties and complex application scenarios make it difficult to detect steel defects, which raises the cost of manual screening. The currently used object detection approach may efficiently find flaws, considerably increase production efficiency, and quicken the transition to an intelligent, modern industry.

Deep learning, as an extension of traditional machine learning, has developed rapidly in recent years in the context of big data. The essence of deep learning is the learning process that enables machines to reach or even surpass human levels. Its unique advantage is that excellent features can be extracted using convolutional networks. Currently, it is widely used in machine vision, pattern recognition, and other fields. A large number of improved algorithms have achieved significant success in terms of accuracy and speed, such as SPPnet, Fast R-CNN, Faster R-CNN, single-shot detector (SSD) [2], You Only Look Once (YOLO), YOLOv2, YOLOv3, YOLOv4, YOLOv5, and other object detection networks. However, it is extremely difficult to achieve a mutual trade-off between speed and precision. So in this work, inspired by YOLO and SSD, we propose an improved mobile-friendly and high-accuracy object detection algorithm. To summarize, our main contributions are as follows:

- We propose an improved YOLOv5 algorithm named MSS-YOLOv5 to improve accuracy while keeping the speed largely unchanged based on YOLOv5 [3–5]. We design an upsampling and downsampling to the network to facilitate deeper information fusion and compensate for missing information. Our design of four YOLO detectors will facilitate the detection of obstacles of different sizes.
- A new pooling method is adopted in the SPP module to improve network performance in this paper. Our pool approach helps reduce information loss compared to maximum pooling and average pooling. This lossless boost will not come at any additional cost to the network. It is friendly to server devices and embedded deployments.
- Inspired by the structure of the SIoU loss function, we add the Angle cost to our loss function. Meanwhile, based on the idea of Focal loss [6], we added the new weight coefficient to the cross-entropy loss function as a way to describe the importance of edge loss to the overall loss function.
- Our improved approach not only performs well on small models but also on large models as well. Referring to the model design of YOLOv5, we present four versions of the model in this paper, MSS-YOLOv5s, MSS-YOLOv5m, MSS-YOLOv5l, and MSS-YOLOv5x.

The rest of this paper is organized as follows. Section 2 introduces the related works. The methods are presented in Section 3. The experiments and results are discussed in Section 4. The conclusions are drawn in Section 5.

2 Related works

With the rise of a deep network, the accuracy of object detection has been greatly improved. The commonly used object detection algorithms are divided into two categories. Two of the most commonly used are two-tier target-detection algorithms that contain regional recommendation networks, such as R-CNN [7], Faster-RCNN [8], Mask-RCNN [9], SPP Net [10], etc. Despite having high accuracy in most detection tasks, these network models have a large number of parameters. They are difficult to deploy on embedded devices and do not have a high recognition accuracy for small targets. The other category is the single-stage YOLO [11–13] (You Only Look Once) family of algorithms. In contrast to the two-stage algorithm, it has a fast inference speed. Because of its ease of deployment, YOLO has a wide range of applications in many areas such as unmanned vehicles and the military.

YOLOv5 is the fifth generation version of YOLO which shows excellent performance in different detect tasks. There are four types of YOLOv5, which are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, while the basic structure of YOLOv5s is shown in Figure 1. Due to the rapid development of deep learning, a large number of excellent works to improve YOLOv5 have emerged. Cheng et al [14] proposed adding attention mechanisms to YOLOv5 to enable the network to learn the information we need adaptively. Xing [15] et al. used YOLOv5 algorithm and DeepSORT algorithm to detect and track multiple moving targets. Lan et al [16]. proposed an improved deep learning network model YOLOv5-DN based on YOLOv5. The CSP-DarkNet module in YOLOv5 was replaced by CSP-DenseNet to promote the accuracy of target detection and classification in the model. Howard et al [17] proposed to combine LRM and Focal loss in YOLOv5 to improve the average accuracy. Zhao et al [18] used the ghost module to reduce the parameters and thus further improve the detection speed. A series of valuable works have contributed to the development of YOLO algorithm.

YOLOv5s was one of the first networks to use SPP in a single-stage algorithm. Although the backbone network can extract some of the features after all the network depth was limited to extract depth information of the network. The SPP module contains convolutional kernels of sizes 1, 5, 9, and 13, which are used to obtain feature information under different perceptual fields by maximum pooling. Although the ASPP [17] and RFB [19] modules have appeared in previous studies, these modules expand the receptive field by dilated convolution and do not

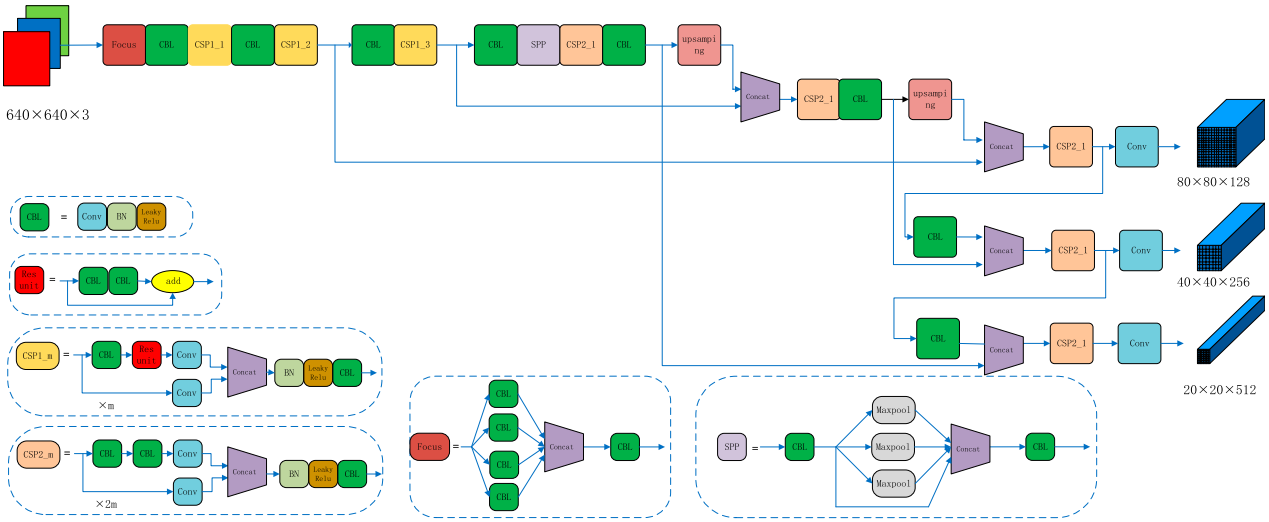


FIGURE 1
The network structure of YOLOv5s. The input size of image is 640 × 640. The final output is three effective feature layers, and their output size is 128, 256, 512. The structure inside the algorithm can be used for the extraction of feature information.

address the information loss problem caused by maximum pooling or average pooling.

There are many factors that affect YOLOv5s performance, such as loss function, backbone networks, pool method, etc. A great many works were emerged to improve the loss function. Li et al. proposed GIOU [20] to solve the problem of disappearing gradients. However, there are some problems such as slow convergence. Zheng et al. On this basis, the DIoU [21] was proposed, and the distance between the mass of the prediction frame and the real frame is considered in the function definition. Cai et al. found that there is a risk of degradation when the centroids of two boxes overlap. The aspect ratio of the boxes was therefore introduced to form the CIoU [22]. Although CIoU considers the overlap area, centroid distance, and aspect ratio, the true difference between aspect and confidence is not well reflected by v in the formula, making it more difficult to optimize. Min et al. then reconsidered the aspect factors and proposed EIoU [23] on top of this. The above work is useful for portraying the difference between the prediction frame and the true frame. There is still room for improvement in the loss function.

The pooling method affects the detection performance of the model to some extent. Kumar et al [24] used a deep network model using ResNet-50 and global average pooling to solve the vanishing gradient and overfitting problems. Tan et al [25] proposed to incorporate maximum pooling into an improved SPP network to enhance the network's ability to represent information. Zhang et al [26] proposed to replace max pooling and average pooling with random pooling to obtain deep learning models with better performance. However, the problem of maximum pooling and average pooling leading to significant information loss has not been resolved.

In summary, the ability of YOLOv5 to extract detailed information is limited, and the balance between speed and accuracy has been a difficult problem to be tackled. YOLOv5 has significant room for improvement, both in terms of the loss function and pooling methods or feature fusion. Therefore, we will also focus on these three improvement points in this paper.

3 Presented network

In this section, we present some of our design ideas about MSS-YOLOv5, which help us trade off between speed and precision. First, we design four branches to integrate different scale features. Then, we replace Maxpool with an improved SoftPool in the SPP module. Finally, inspired by the structure of the SIoU loss function, we add the Angle cost and other strategies to improve the performance further.

Combining these approaches, we named the improved YOLOv5 algorithm MSS-YOLOv5. MSS takes the initials multi-scale fusion, Softpool, and SIoU respectively. Similar to YOLOv5, we provide four versions, with the number of model parameters ranging from small to large as MSS-YOLOv5s, MSS-YOLOv5m, MSS-YOLOv5l, and MSS-YOLOv5x. The overall flow of the model is shown in Figure 2. After the data is enhanced, the input pictures are sent into the model for training. BCE loss is used to calculate the classification loss and target loss. NMS (non-maximum

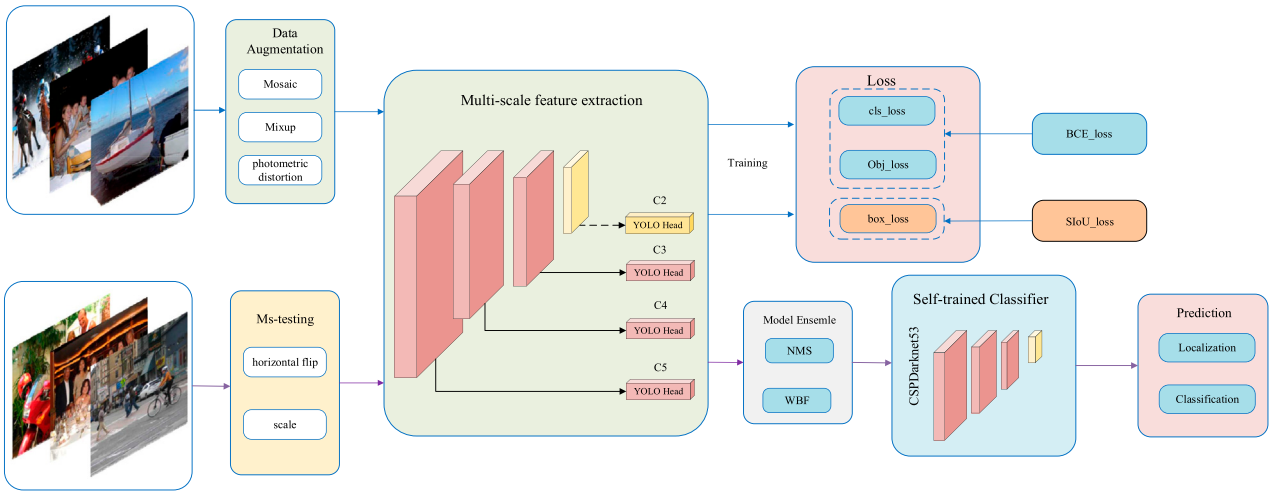


FIGURE 2
The process of training and testing the model. Data augmentation is used to acquire high-quality pictures. SIoU loss can help us reduce the loss of border to train the model more effectively. The model ensemble contains different methods, which can remove those detection boxes with a low score. The prediction section consists of two parts: location and classification.

suppression) is used to filter out the boxes with low scores due to occlusion and other factors. Firstly, the IoU threshold is set to 0.5. Secondly, all the boxes are sorted, and each box with IoU >0.5 is set to 0 if it has the highest probability of scoring, and the opposite is kept. The final output is the location and labels information of the target.

3.1 Multi-scale feature integration

There are many large differences in the size and shape of targets in detection tasks. To address this problem, both Scaled-YOLOv4 [27] and TPH-YOLOv5 [28] use a multi-scale feature fusion strategy [29, 30] to extract more useful information. Both decrease the difficulty of detecting target objects of different sizes by increasing the number of detectors. Inspired by these two algorithms, we add a branch to integrate different channel information in the network. And we can use the Concat operation to integrate these information. A large detector has extensive coverage and abundant information on feature points, so it is easier to obtain global information. On the contrary, the small-scale detector has opposite characteristics. The loss of feature information is more obvious after the backbone network. Generally, only single-digit pixel sizes are left, resulting in small targets that do not match the original image after mapping through the perceptual field, which ultimately leads to poor regression and prediction results. Backbone, SPP, and PANet in YOLOv5s have extracted some feature information about the target to a certain extent. However, there is still some room for mining deep semantic information and shallow detailed information.

According to the above problems, we proposed the following improved measures. 1) As shown in Figure 3, we add one more upsampling and downsampling in the PANet of YOLOv5s (PANet originally had two upsamples and two downsamples). The sampling structure consists of Convolution, Batch Normalization [31], and Leaky Relu [32]. Convolution is used for feature extraction. Batch Normalization can prevent gradients from exploding or disappearing, speed up network convergence and improve the stability of the detection network. Leaky Relu can enhance the ability of non-linear representation of the network. 2) We add an extra YOLO Head as a detector to accommodate different scales of target detection. The multi-scale fusion strategies used in this paper are all methods of fusion at four different scales.

3.2 Improved SoftPool

The main function of pooling is to reduce the dimensionality of the feature map, reducing the computational overhead and thus saving memory, offering the possibility of studying deeper networks. The prevailing pooling methods are maximum pooling and average pooling or a combination of both, but extensive experiments have shown that these types of pooling result in the loss of important feature information. Therefore, literature 32 proposes the SoftPool [33] method, where each activation is assigned a corresponding weight through a softmax operation. The weights can be expressed as follows.

$$\omega_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}} \tag{1}$$

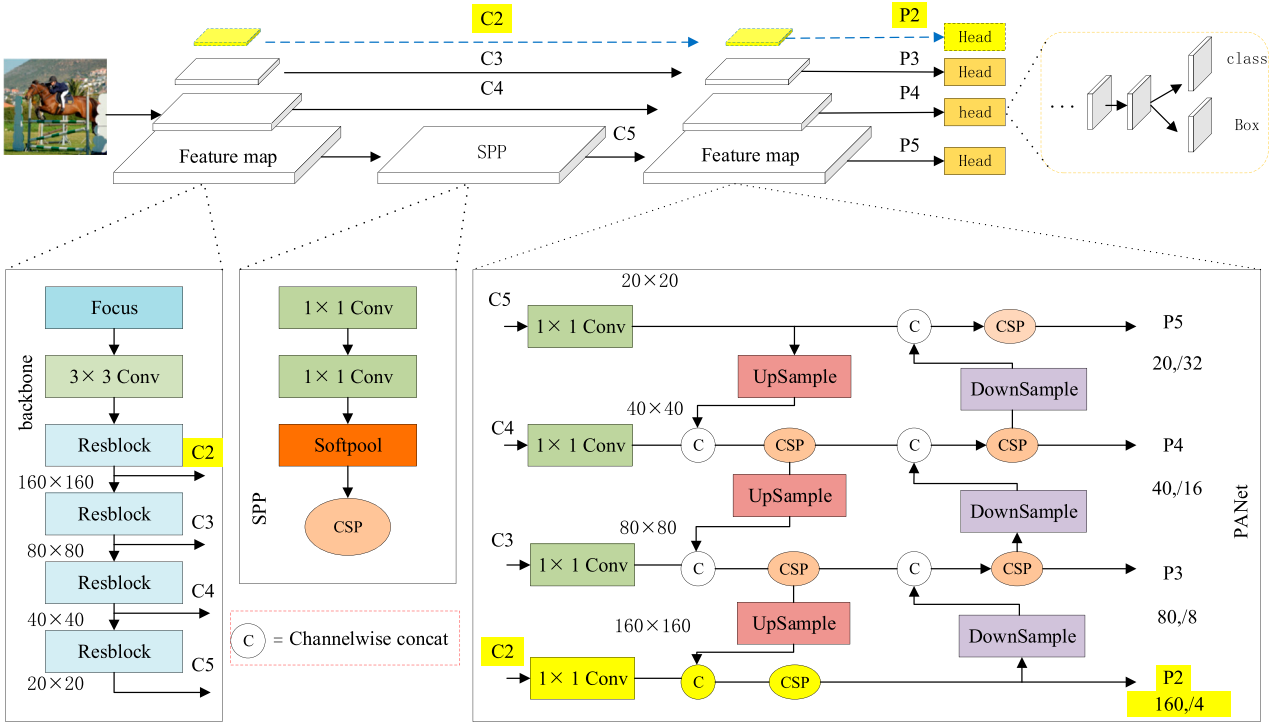


FIGURE 3
Multi-scale feature integration architecture. The Backbone is CSPDarknet53, which outputs C2–C5 feature maps to the neck. The neck is a new PANet, which inputs four feature maps and outputs four feature maps. For YOLOv5-MSSs, the input channel numbers are [20, 40, 80], and the output channel numbers are [20, 40, 80, 160].

The output of soft pooling ($\bar{a}$) is the weighted sum of all activations in the kernel neighborhood R.

$$\bar{a} = \sum_{i \in R} \omega_i \cdot a_i \tag{2}$$

Soft pooling performs a normalization operation using the softmax of a region. Its probability distribution is proportional to each activation value relative to the neighboring activation values in the kernel region. Therefore, SoftPool is microscopic. It can provide a certain gradient at each backpropagation. However, there are still problems such as limited lifting accuracy and the return value of the gradient is too small to be optimized. Therefore, a new SPP structure is proposed in this paper. As shown in Figure 3, the MaxPool in the SPP is replaced with SoftPool, while the pooling kernel size is adjusted from [5, 9, 13] to [3, 5, 7] to retain rich enough depth information and enhance feature representation. Of course, it is possible to keep the convolution kernel size the same or resize it to [5, 7, 9]. However, kernel sizes of [3, 5, 7] are significantly less computationally intensive. At the same time, when the fitting ability of the network is saturated, it will be beneficial to reduce more redundant information.

3.3 SIoU loss

Object detection is one of the core problems in the field of vision and its detection accuracy depends on the definition of the loss function. In previous studies, the loss function has mostly been defined using the distance, intersection ratio, and aspect ratio between the prediction box and the true box. We have not taken into account the direction in which the predicted boxes do not match the real boxes. The loss function has disadvantages such as slow convergence, difficulty in optimization, and low detection accuracy. Therefore, we adopt a new loss function SIoU in this paper. SIoU was pioneered by Zhora Gevorgyan [34] in 2022 and consists of four main Cost functions, Angle cost, Distance cost, Shape cost, and IoU cost. The latter three elements have been studied enough in previous work to have a positive impact. However, it does not mean that there is no room for improvement in the loss function. So Angle cost is added. This addition ensures that the prediction is effective. This improved method allows the prediction box to be moved quickly to the nearest axis. Finally, only the X or Y coordinates are needed for the regression operation. Overall, the Angle cost penalty makes the degrees of freedom of loss much lower, making it

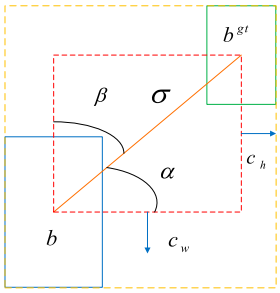


FIGURE 4
The relationship between the position of the prediction box and the real box.

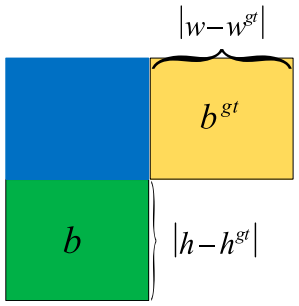


FIGURE 5
Schematic of relation of IoU component contribution.

easier to converge. The following sections show the computation of the four Cost functions.

3.3.1 Angle cost

The picture of regression loss of borders was shown in Figure 5. It reflects the relationship between the position of the predicted box and the real box. We calculate the relevant parameters in Figure 4.

In order to make the function converge quickly, we will first try to minimize α if $\alpha \leq \frac{\pi}{4}$ otherwise minimize $\beta = \frac{\pi}{2} - \alpha$.

To achieve this first, an angle-aware component is introduced and defined as follows:

$$\Lambda = 1 - 2 \cdot \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{3}$$

Where

$$x = \frac{c_h}{\sigma} = \sin(\alpha) \tag{4}$$

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{cy}\right)^2} \tag{5}$$

$$c_h = \max\left(b_{c_y}^{gt} - b_{c_y}\right) - \min\left(b_{c_y}^{gt} - b_{c_y}\right) \tag{6}$$

b and b^{gt} are the centers of the predicted and real boxes respectively. σ is the distance between the center point of the predicted box and the real box. c_w and c_h denote the width and height of the rectangle with σ as the diagonal, respectively. α and β denote the angles formed by the diagonal and the width and height respectively, of which $\alpha + \beta = \frac{\pi}{2}$.

3.3.2 Distance cost

The distance is defined in the following way:

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma \rho_t}\right) \tag{7}$$

$$\rho_x = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_w}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda \tag{8}$$

The contribution of Distance cost is small when the angle is small but becomes larger as the angle gradually converges to $\frac{\pi}{4}$.

3.3.3 Shape cost

The shape is defined in the following way:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega_t}\right)^\theta \tag{9}$$

where

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{10}$$

θ reflects the degree of attention paid to Shape cost and θ is uniquely determined for each dataset. $\theta = 4$ is calculated by the genetic algorithm in this paper.

3.3.4 IoU cost

IoU [35] reacts to the ratio of intersection to concatenation when the prediction box intersects the real box. A Schematic of the relation of IoU component contribution was shown in Figure 5. The formula is as follows.

$$IoU = \frac{|b \cap b^{gt}|}{|b \cap b^{gt}|} \tag{11}$$

3.3.5 SIoU cost

The regression loss of the border is represented below.

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{12}$$

3.3.6 Total loss

The final loss function used in this paper is as follows.

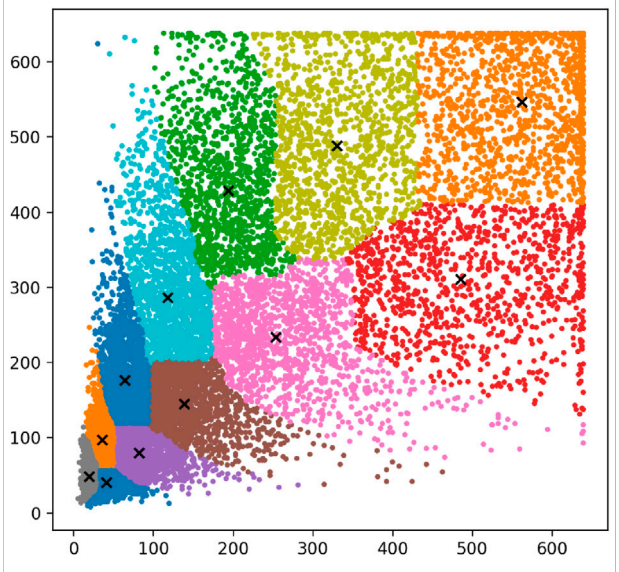


FIGURE 6
Map of clustering centers.

FIGURE 7
The results of the mosaic data enhancement.

$$
\begin{aligned}
Loss &= \alpha Loss_{SIoU} + \beta Loss_{conf} + \gamma Loss_{cls} \\
&= \alpha \cdot \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(x_i - \hat{x}_i^j \right)^2 + \left(y_i - \hat{y}_i^j \right)^2 \right] \\
&\quad + \alpha \cdot \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\sqrt{\omega_i^j} - \sqrt{\hat{\omega}_i^j} \right)^2 + \sqrt{h_i^j} - \sqrt{\hat{h}_i^j} \right)^2 \Bigg] \\
&\quad - \beta \cdot \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\hat{C}_i^j \log\left(C_i^j \right) + \left(1 - \hat{C}_i^j \right) \log\left(1 - C_i^j \right) \right] \\
&\quad - \beta \cdot \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} \left[\hat{C}_i^j \log\left(C_i^j \right) + \left(1 - \hat{C}_i^j \right) \log\left(1 - C_i^j \right) \right] \\
&\quad - \gamma \cdot \sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in classes} \left[\hat{P}_i^j \log\left(P_i^j \right) + \left(1 - \hat{P}_i^j \right) \log\left(1 - P_i^j \right) \right]
\end{aligned} \tag{13}
$$

In all the above formulas, $Loss_{SIoU}$ means SIoU Loss, $Loss_{conf}$ means confidence loss, $Loss_{cls}$ means class Loss. α, β, γ denote the weighting factors respectively, which are used to measure the importance of different losses. In this paper, α, β and γ take the values 0.5, 1, and 4 respectively.

3.4 Other strategies

The K-means [36] clustering method was chosen to predict more accurate anchor frames in this paper. Different types of objects have different sized frames, and the same object may vary depending on how close or far it is photographed. Each detector uses three anchor frames to determine the position of the object. There are three detectors in YOLOv5s, so nine clustering centers are needed. Based on the experiments conducted, the final clusters were: [(19,48), (40,40), (35,97), (81,80), (64,176), (138,145), (117,286), (253,234), (193,428), (485,310), (330,488), (561,546)]. The distribution of clustering centers is shown in Figure 6.

Data augmentation is a common way of expanding data. It can enhance the detection capability of a neural network with a limited amount of data. In this paper, we have the requirement to enhance the generalization capability of the model. Therefore, we adopted the mosaic data augmentation method to stimulate the maximum performance of the algorithm. In previous enhancement methods, horizontal inversion and illumination were often used to enhance the data, but there were many drawbacks such as poor generalization and hindering accuracy improvement. Therefore, we follow the mosaic data enhancement method [37] of YOLOv4 and YOLOX [38] in this paper. Numerous experiments have shown that this enhancement method has an effort on increasing the detection accuracy and enhancing the generalization ability of the model to a certain extent. The results of the mosaic data enhancement are shown in Figure 7.

4 Experiments and results

4.1 Experimental environment and datasets sources

The hardware setup in the laboratory configured for this study is as follows: the experimental platform is Windows 10, the processor is Intel Core i7-11700F 2.50 GHZ, equipped with NVIDIA GeForce RTX3060-32GB, the development environment is Pycharm2020, Python3.6, the deep learning framework is Pytorch1.7, using CUDA11.2.0/CUDNN11.2 for image acceleration.

The public PASCAL VOC datasets used in the training process are as follows:

1) PASCAL VOC 2007: a real-world dataset with still different views from our life. It contains 20 categories with a total of 4952 pictures. Moreover, the training and test sets were divided according to 9:1, with 4457 training sets and 495 test sets.
2) PASCAL VOC 2012: a real-world dataset with still different views from our life. It contains 20 categories with a total of 17125 pictures. Moreover, the training and test sets were divided according to 9:1, with 15412 training sets and 1713 test sets.

The dataset of PASCAL VOC 2007 and PASCAL VOC 2012 were used to validate the effort of the improved method. At the same time, We compared common lightweight networks for comparative experiments. FPS and mAP were combined to compare the superiority of the algorithms.

4.2 Evaluation indicators

Precision, recall, AP, and mAP are used to evaluate the merits of the model. The formulae are shown below.

$$P_{precision} = \frac{TP}{TP + FP} \tag{14}$$

$$R_{recall} = \frac{TP}{TP + FN} \tag{15}$$

$$AP = \int_0^1 P(R)dR \tag{16}$$

$$mAP = \frac{1}{C}\sum_{c \in C} AP(c) \tag{17}$$

TP represents the total number of correctly classified positive samples, FP represents the total number of misclassified positive samples and FN represents the total number of misclassified negative samples. The precision rate indicates the number of positive category samples as a proportion of the total number of samples. The recall indicates the proportion of all positive samples detected to the number of positive samples in the dataset. The mAP can be used as a comprehensive evaluation metric for single category detection, with higher AP values indicating better detection of a category, and mAP being a comprehensive evaluation of the entire network. The complexity of a model is measured by the number of parameters or computations. In general, the lower the number of parameters in a model, the faster the detection speed, which is usually evaluated in terms of FPS.

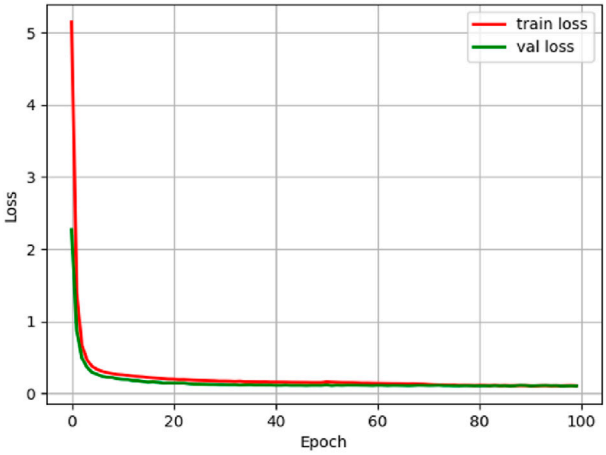


FIGURE 8
Loss function curve of MSS-YOLOv5s.

4.3 Model training

The following settings are made when the model is trained. To obtain better training results, this experiment uses the pre-training weights of the CSPDarknet53 backbone, and the model is optimally trained using SGD (stochastic gradient descent). The input image size of the model is 608 × 608, the maximum learning rate is 1e−2, the freeze part batch size is 16, freeze training for 50 cycles, unfreeze part batch size is 50 cycles of freeze training. The IoU threshold and momentum are set to 0.5 and 0.937, respectively. Other versions of MSS-YOLOv5 use the same training method.

The model was trained using the above parameter settings and a combination of improvements. The final loss function curve is shown in Figure 8. From the figure, we can see that the loss function curve has an overall decreasing trend. Meanwhile, the loss curve has almost approached convergence at the 20th epoch. The experiments demonstrate that our method is not only easy to converge but also highly stable.

TABLE 1 Ablation experiment of MSS-YOLOv5s.

Multi-scale	Softpool	SIoU	mAP/%	FPS/f/s	Model size/MB
✗	✗	✗	81.49	65	27.14
✓	✗	✗	82.03	54	27.70
✓	✓	✗	82.73	51	27.70
✓	✓	✓	**84.39**	50	27.70

The bolded values indicate the best experimental results in the same group of experiments.

TABLE 2 Comparison of different algorithms.

Model	Backbone	mAP/%	FPS/f/s	Model size/MB
YOLOv3	Darknet53	79.68	37	235.08
YOLOv4	CSPDarknet53	85.23	24	248.25
YOLOv4- tiny	CSPDarknet53-Tiny	77.47	116	22.58
Faster RCNN	Resnet50	77.42	7	522.91
YOLOv5s	CSPDarknet53	81.49	65	27.14
YOLOv5m	CSPDarknet53	87.73	15	80.62
YOLOv5l	CSPDarknet53	90.81	10	176.39
YOLOv5x	CSPDarknet53	92.67	6	329.38
MSS-YOLOv5s(Ours)	CSPDarknet53	**84.39**	50	27.70
MSS-YOLOv5m(Ours)	CSPDarknet53	89.11	14	82.31
MSS-YOLOv5l(Ours)	CSPDarknet53	91.53	9	182.05
MSS-YOLOv5x(Ours)	CSPDarknet53	**92.87**	6	340.04

The bolded values indicate the best experimental results in the same group of experiments.

4.4 Ablation experiment

To verify the effectiveness of the algorithm, we conducted ablation experiments on the improved modules, in order of four scales, maximum pooling replacement to Softpool, and GIoU replacement to SIoU, to verify the detection effectiveness of the improved algorithm. Through experiments, we found that the improved method has significant performance gains on small models, but not much for large models. Therefore, we demonstrate ablation experiments with MSS-YOLOv5s as an example in this paper. The results of the ablation experiments of MSS-YOLOv5s are shown in Table 1.

As we can see that quadruple scale feature fusion, Softpool, and SIoU loss function, provide a significant improvement in detection accuracy from Table 2. The multi-scale fusion sacrifices some of the speed, but after all, it is minimal and gives a solution for accuracy improvement. With the introduction of softpool and SIoU, the model size remains almost unchanged and the speed is essentially the same, with an average precision improvement of 2.9%.

4.5 Comparison of different algorithms

To reflect the effectiveness of the algorithm improvements, we experimentally compared the target detection algorithms YOLOv4, YOLOv4-tiny, YOLOv3, YOLOv3-tiny, and YOLOv5. The experimental results are shown in Table 2.

From Table 3, we can see that although the two-stage Faster RCNN uses a region suggestion network, it does not achieve higher accuracy. On the contrary, YOLOv4 works better but poses some difficulties for model deployment due to its slower speed. YOLOv4-tiny, YOLOv3-tiny, and YOLO5s, as commonly used lightweight algorithms, have certain advantages, but the detection accuracy is too low to meet the needs of autonomous driving corresponding to complex scenarios. In the improved model, MSS-YOLOv5s, MSS-

TABLE 3 Performance of different algorithms on the PASCAL 2012 dataset.

Model	Backbone	mAP/%	FPS/f/s	Model size/MB
YOLOv3	Darknet53	79.88	28	235.08
YOLOv4	CSPDarknet53	85.49	21	248.25
YOLOv4- tiny	CSPDarknet53-Tiny	77.52	111	22.58
Faster RCNN	Resnet50	77.81	6	522.91
YOLOv5s	CSPDarknet53	82.04	62	27.14
YOLOv5m	CSPDarknet53	87.81	15	80.62
YOLOv5l	CSPDarknet53	90.83	12	176.39
YOLOv5x	CSPDarknet53	92.68	6	329.38
MSS-YOLOv5s(Ours)	CSPDarknet53	84.44	49	27.70
MSS-YOLOv5m(Ours)	CSPDarknet53	89.17	14	82.31
MSS-YOLOv5l(Ours)	CSPDarknet53	91.04	11	182.05
MSS-YOLOv5x(Ours)	CSPDarknet53	92.91	6	340.04

Through the above comparison, we can easily find that MSS-YOLOv5 not only maintains a faster speed but also outperforms other lightweight networks in terms of accuracy. It proves that MSS-YOLOv5 can work effectively on different datasets.

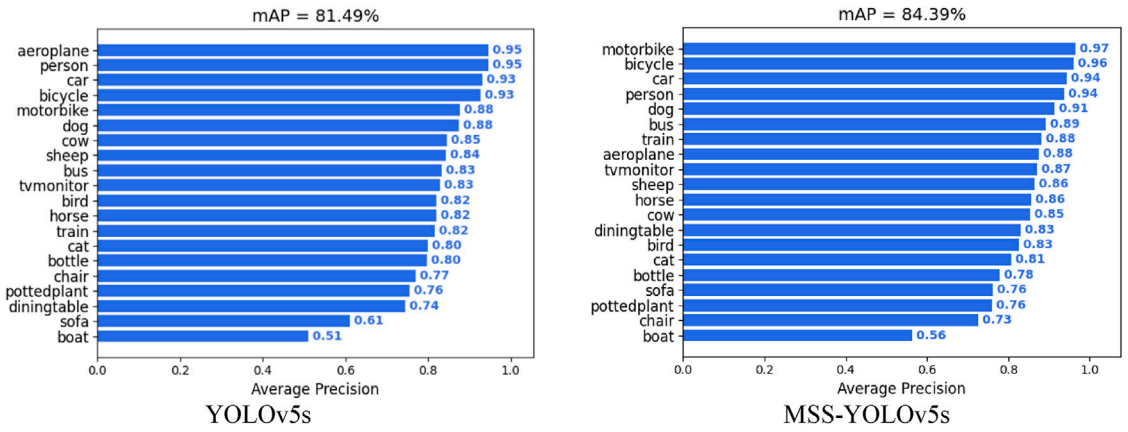


FIGURE 9
Comparison of the mAP of YOLOv5s and MSS-YOLOv5s.

YOLOv5m, MSS-YOLOv5l, and MSS-YOLOv5x have different degrees of enhancement. The speed of MSS-YOLOv5s is essentially the same as YOLOv5s, but there is a significant improvement in mAP. This is despite a 0.2% improvement in the large model MSS-YOLOv5x, which tends to be saturated. This non-destructive improvement of MSS-YOLOv5 is extremely model friendly, achieving a degree of balance between speed and accuracy and providing more options for embedded deployments.

From Figure 9, we can see that the detection accuracy of all types of targets has been improved to different degrees, especially for small targets. When using YOLOv5s, the detection effect of the dining table, sofa and boat are not obvious, but on our improved algorithm, the improvement is 0.09%, 0.15%, and 0.05% respectively, which shows that our improved strategy is simple and effective.

To give a more intuitive picture of the detection effect of the improved algorithm on the PASCAL VOC2007 dataset, Figure 10 shows the detection of the different algorithms, the right panel shows the detection effect of the original YOLOv5s and YOLOv5x model and the left panel shows the detection effect of the MSS-YOLOv5x and MSS-

FIGURE 10
Comparison of test results for different algorithms.

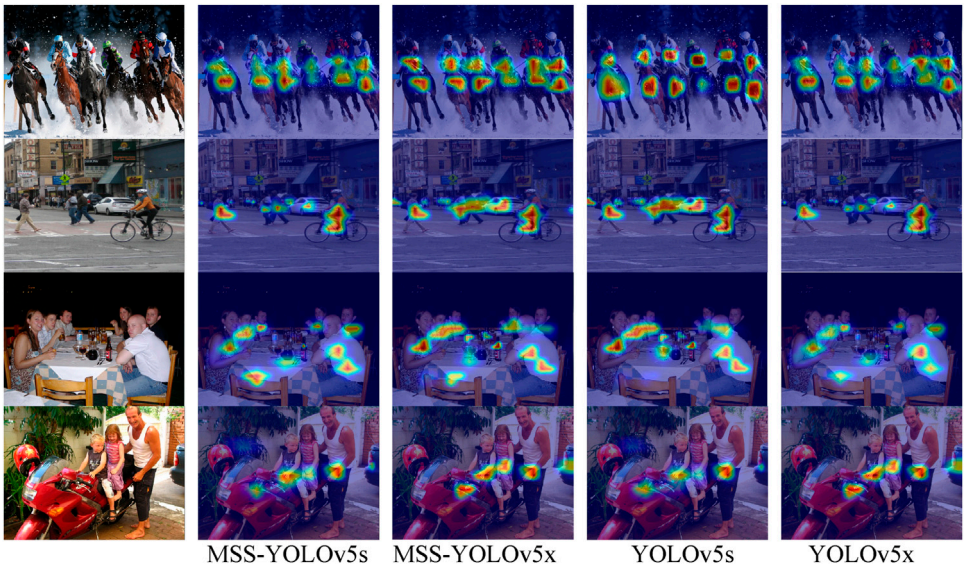


FIGURE 11
Comparison of different algorithms for heat maps.

YOLOv5s. From the figure, we can see that the MSS-YOLOv5x detected significantly more targets than the YOLOv5x algorithm, and for the targets that were both detected, the confidence level of the MSS-YOLOv5x was higher. The same result is found on the MSS-YOLOv5s and YOLOv5s. This shows that our improved approach improves the performance of the model both on large and small models. The MSS-YOLOv5 not only enriched the deep semantic information of the feature map but also enhanced the acquisition of shallow detail information to a certain extent, improving the detection capability of the network for targets of different sizes.

4.6 Performance on the PASCAL VOC2012 dataset

To further validate the effectiveness of the improved method, we used the PASCAL VOC 2012 dataset to prove the superior performance of the new framework. The same training approach was used to retrain the PASCAL VOC 2012 dataset. Figure 11 is a heat map presentation of the different algorithms on the PASCAL VOC 2012 dataset.

The heat map represents the area of interest of the network to the detection target, and the more thermal points, the more targets are detected. Experiments have shown that our algorithm is still able to obtain better detection results.

The performance of the different algorithms on PASCAL 2012 is shown in Table 3.

5 Conclusion

In this work, we propose an improved YOLOv5 object detection algorithm named MSS-YOLOv5 to solve the problem of a trade-off between the speed and precision of YOLOv5 in object detection. Multi-scale information is integrated into different feature dimensions to improve the distinction and robustness of features. The design of the detectors increases the variety of detection boxes to accommodate a wider range of detected objects. The pooling method is upgraded to obtain more detailed information. We add the Angle cost and assign new weights to different loss functions to accelerate the convergence and improve the accuracy of network detection. Experiments have shown that the improved model has essentially similar inference speeds to the original model. However, the improvements we propose are effective in improving accuracy on both large and small models and perform well on different data sets. SIoU loss and feature fusion approaches can be considered to optimize other network structures. We propose a new model with reliable accuracy and high timeliness.

The presented network not only achieves great performance on the PASCAL 2007 but also works efficiently on the PASCAL 2012 dataset. However, our proposed more efficient deep learning-based YOLO series algorithm still cannot work perfectly to heavily obscured targets. In the future, we will introduce structural reparameterization techniques in backbone and FPN to improve the overall performance of your network and add swin transformerv2 to backbone to enhance the network's ability to capture information over long distances.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://pjreddie.com/projects/pascal-voc-dataset-mirror/.

Author contributions

Conceptualization, YH and YS; methodology, YH and YL; software, YH; validation, YH, YS, and JY; formal analysis, XW; writing—original draft preparation, YH; writing—review and editing, YS; supervision, XW and YL. All authors have read and agreed to the published version of the manuscript.

Funding

This work is sponsored by the Natural Science Foundation of Chongqing (Grant No. cstc2019jcyj-msxmX0220, No. CSTB2022NSCQ-MSX1425, CSTB2022NSCQ-MSX0398), Science and Technology Foundation of the Education Department of Chongqing (Grant No. KJQN202101510), China.

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 T, Tang M, Li H, Li J, Zou Y, Pan Y, et al. A multidirectional pendulum kinetic energy harvester based on homopolar repulsion for low-power sensors in new energy driverless buses. *Int J Precision Eng Manufacturing-Green Technol* (2022) 9(2):603–18. doi:10.1007/s40684-021-00344-5

2. Zheng W, Tang W, Jiang L. SE-SSD: Self-ensembling single-stage object detector from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, June 20–25, 2021 (2021). p. 14494.

3. Hu GX, Hu BL, Yang Z, Huang L, Li P. Pavement crack detection method based on deep learning models. *Wirel Commun Mob Comput* (2021) 2021(1):1–13. doi:10.1155/2021/5573590

4. Wu TH, Wang TW, Liu YQ. Real-time vehicle and distance detection based on improved yolo v5 network[C]. In: 2021 3rd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, June 18–20, 2021. *IEEE* (2021) 24.

5. Ting L, Baijun Z, Yongsheng Z. Ship detection algorithm based on improved YOLO V5. In: 2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE), Guangzhou, China, July 15–17, 2021 (2021). p. 483.

6. Li X, Wang W, Hu X. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021). p. 11632–41.

7. Girshick R. Fast R-CNN[C]. In: IEEE International Conference on Computer Vision (ICCV), December 7–13, 2015, Santiago, Chile (2015). p. 1440–8.

8. Ren S, He K, Girshick R, Sun J. 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

9. He K, Gkioxari G, Dollár P, Girshick R. Mask RCNN[C]. In: IEEE International Conference on Co-mputer Vision (ICCV), October 22–29, 2017, Venice, Italy (2017). p. 2980.

10. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans pattern Anal machine intelligence* (2015) 37(9):1904–16. doi:10.1109/tpami.2015.2389824

11. Wang X, Wang S, Cao J, Wang Y. Data-driven based tiny-YOLOv3 method for front vehicle detection inducing SPP-Net. *IEEE Access* (2020) 8(99):110227–36. doi:10.1109/access.2020.3001279

12. Redmon J, Farhadi A. YOLO9000:Better, faster, stronger[C]. In: IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, July 21–26 2017 (2017). p. 6517.

13. Redmon J, Farhadi A. Yolov3: An incremental improvement[C]. In: IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, June 18–23, 2018 (2018). arXiv: 1804.0276.

14. Cheng L, Li J, Duan P, Wang M. A small attentional YOLO model for landslide detection from satellite remote sensing images. *Landslides* (2021) 18(8): 2751–65. doi:10.1007/s10346-021-01694-6

15. Xing C, Sun B, Zhang W. Image-enhanced YOLOv5 and deep sort underwater multi-moving target tracking method[C]. In: 2022 5th international symposium on autonomous systems (ISAS), Hangzhou, China, April 08–10, 2022. IEEE (2022) 1–6.

16. Lan Y, Xu W. Insulator defect detection algorithm based on a lightweight network. *J Phys Conf Ser* (2022) 2181(1):012007. doi:10.1088/1742-6596/2181/1/012007

17. Howard A, Sandler M, Chen B. Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), October 27–November 2, 2019. IEEE (2020).

18. Zhao T, Wei X, Yang X. Improved YOLO v5 for railway PCCS tiny defect detection[C]//2022 14th international conference on advanced computational intelligence (ICACI). *IEEE* (2022) 85–90.

19. Zhang W, Wang J, Guo X, Chen K. Two-stream RGB-D human detection algorithm based on RFB network. *IEEE Access* (2020) 8(99):123175–81. doi:10.1109/access.2020.3007611

20. XiaoFan L, HaiBo P, Yi W, JiangChuan L. "Introduce GIoU into RFB net to optimize object detection bounding box[C]," in Proceedings of the 5th International Conference on Communication and Information Processing, Chongqing, China, November 15–17, 2019 (2019), 108–113.

21. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc AAAI Conf Artif Intelligence* (2020) 34(07):12993–3000. doi:10.1609/aaai.v34i07.6999

22. Cai Z, VasconcelosCascade NR-CNN. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans pattern Anal machine intelligence* (2019) 43(5):1483–98. doi:10.1109/tpami.2019.2956516

23. Min Y, Guo J, Yang K. Research on real-time detection algorithm of rail-surface defects based on improved YOLOX[J]. *J Appl Sci Eng* (2022) 26(6):801. doi:10.6180/jase.202306_26(6).0006

24. Kumar RL, Kakarla J, Isunuri BV, Singh M. Multi-class brain tumor classification using residual network and global average pooling. *Multimedia Tools Appl* (2021) 80(9):13429–38. doi:10.1007/s11042-020-10335-4

25. Tan YS, Lim KM, Tee C, Lee CP, Low CY. Convolutional neural network with spatial pyramid pooling for hand gesture recognition[J]. *Neural Comput Appl* (2021) 33(10):5339–51. doi:10.1007/s00521-020-05337-0

26. Zhang YD, Satapathy SC, Liu S, Li GR. A five-layer deep convolutional neural network with stochastic pooling for chest CT-based COVID-19 diagnosis. *Machine Vis Appl* (2021) 32(1):14–3. doi:10.1007/s00138-020-01128-8

27. Wang CY, Bochkovskiy A, Liao HYM. Scaled-yolov4: Scaling cross stage partial network[C]. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, Nashville, TN, June 20–25, 2021 (2021). p. 13029.

28. Zhu X, Lyu S, Wang X. "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios [C]," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, October 10–17, 2021 (2021). 2778–88.

29. Li Y, Wang X, Zhu Z. A novel person re-id method based on multi-scale feature fusion[C]. In: 2020 39th Chinese Control Conference (CCC). IEEE (2020). p. 7154–9.

30. Li Y, Xu P, Zhu Z. Real-time driver distraction detection using lightweight convolution neural network with cheap multi-scale features fusion block. In: Proceedings of 2021 Chinese Intelligent Systems Conference, Fuzhou, China, October 16–17, 2021. Singapore: Springer (2022). p. 232.

31. Chen H, Wang YH, Fan CH. A convolutional autoencoder-based approach with batch normalization for energy disaggregation. *J Supercomputing* (2021) 77(3): 2961–78. doi:10.1007/s11227-020-03375-y

32. Dubey SR, Chakraborty S. Average biased ReLU based CNN descriptor for improved face retrieval. *Multimedia Tools Appl* (2021) 80(15):23181–206. doi:10.1007/s11042-020-10269-x

33. Stergiou A, Poppe R, Kalliatakis G. "Refining activation downsampling with SoftPool," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, October 10–17, 2021 (2021), 10357–66.

34. Zheng J, Wu H, Zhang H, Wang Z, Xu W. Insulator-defect detection algorithm based on improved YOLOv7[J]. *Sensors* (2022) 22(22):8801. doi:10.3390/s22228801

35. Ni X, Ma Z, Liu J, Shi B, Liu H. Attention network for rail surface defect detection via consistency of Intersection-over-Union (IoU)-Guided Center-Point Estimation[J]. *IEEE Trans Industr Inform* (2021) 18(3):1694–705. doi:10.1109/TII.2021.3085848

36. Abo-Elnaga Y, Nasr S. K-means cluster interactive algorithm-based evolutionary approach for solving bilevel multi-objective programming problems. *Alexandria Eng J* (2022) 61(1):811–27. doi:10.1016/j.aej.2021.04.098

37. Dawson WO, Kuhn CW. Enhancement of cowpea chlorotic mottle virus biosynthesis and *in vivo* infectivity by 2-thiouracil. *Virology* (1972) 47(1):21–9. doi:10.1016/0042-6822(72)90234-6

38. Zheng H, Wang G, Li X. YOLOX-Dense-CT: A detection algorithm for cherry tomatoes based on YOLOX and DenseNet[J]. *J Food Meas Charact* (2022) 16(6): 4788–99. doi:10.1007/s11694-022-01553-5

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fphy.2022.1115764

OPEN ACCESS

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

REVIEWED BY
Yiwen Chen,
Wuhan University, China
Shuanglin Yan,
Nanjing University of Science and Technology, China
Jinting Zhu,
Massey University, New Zealand
Jian Pang,
China University of Petroleum, China

*CORRESPONDENCE
Xiaofeng Wang,
xfwang828@126.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 04 December 2022
ACCEPTED 29 December 2022
PUBLISHED 12 January 2023

CITATION
Wang X, Sun J, Qin H, Yuan Y, Yu J, Su Y and Sun Z (2023), Accurate unsupervised monocular depth estimation for ill-posed region.
Front. Phys. 10:1115764.
doi: 10.3389/fphy.2022.1115764

Accurate unsupervised monocular depth estimation for ill-posed region

Xiaofeng Wang[1]*, Jiameng Sun[2], Hao Qin[2], Yuxing Yuan[1], Jun Yu[2], Yingying Su[2] and Zhiheng Sun[2]

[1]College of Mathematical and Physical Sciences, Chongqing University of Science and Technology, Chongqing, China, [2]College of Electrical Engineering, Chongqing University of Science and Technology, Chongqing, China

Unsupervised monocular depth estimation is challenging in ill-posed regions, such as weak texture scenes, projection occlusion, and redundant error of detail information, etc. In this paper, in order to tackle these problems, an improved unsupervised monocular depth estimation method for the ill-posed region is proposed through cascading training depth estimation network and pose estimation network by loss function. Firstly, for the depth estimation network, a feature extraction network using asymmetric convolution is designed instead of traditional convolution, which strengthens the extraction of the feature information and improves the accuracy of the weak texture scenes. Meanwhile, a feature extraction network integrating multi-scale receptive fields with the structure of different scale convolution and dilated convolution stack is designed to increase the underlying receptive field of the depth estimation network, which strengthens the fusion ability of the network for multi-scale detail information, and improves the integrity of the model output details. Secondly, a pose estimation network using an attention mechanism is presented to strengthen the pose detail information of keyframes and suppress redundant errors of the pose information of non-keyframes. Finally, a loss function with minimum reprojection error is adopted to alleviate the occlusion problem of the projection process between adjacent pixels and enhance the quality of the output depth images of the model. The experiments demonstrate that our method achieves state-of-the-art performance on KITTI monocular datasets.

KEYWORDS

unsupervised monocular depth estimation, asymmetric convolution, multi-scale receptive field, attention mechanism, ill-posed regions

1 Introduction

As an important research focus in the field of computer vision, monocular depth estimation aims to explore the mapping relationship between image and depth, and predict the depth information from a single image. Monocular depth estimation plays an important role in visual tasks, especially in intelligent fields such as autonomous driving, 3D map construction, AR (Augmented Reality) synthesis, etc.

At present, the mainstream way of monocular depth estimation task is to train the deep neural network by using a large number of marked real depth images as the training set, so as to obtain the depth value of the corresponding pixel from the image. In this way, deep neural networks are used to generate high-quality depth images with different optimization strategies [1–4]. However, supervised depth estimation methods need to collect a large amount of real-depth information data and require an immense amount of computing time in the training

process, which greatly increases the difficulty and complexity of the algorithm. Comparatively speaking, unsupervised monocular depth estimation only requires monocular video sequences or stereo image pairs to realize the depth information estimation of each pixel of a single image [5–7]. In recent years, unsupervised monocular depth estimation have been favored by researchers [8–10]. Among them, Zhou [10] innovatively proposes an unsupervised training framework which cascades the depth estimation network and the pose estimation network through the loss function to predict the depth information of the image, improving the accuracy of model estimation and becoming one of the most dominant frameworks in current unsupervised monocular depth estimation.

However, current unsupervised monocular depth estimation studies, including Zhou's method, still face great challenges in dealing with ill-posed regions problems, such as weak texture scenes, occlusion of pixel projections, and lack of detailed information in depth images, etc. As a result, the depth information obtained by the model cannot fully reflect the image-depth mapping relationship. To solve these problems, we propose an improved unsupervised monocular depth estimation which included a depth estimation network, pose estimation network, and the loss function. Firstly, in the depth estimation network, asymmetric convolution structure and multi-scale field structure are proposed to enhance the feature extraction capability of the network, to alleviate the influence of weak texture scenes. Secondly, in the pose estimation network, the redundant information of pose estimation of adjacent image frames is reduced by the attention mechanism structure. Finally, the minimum reprojection error is introduced into the loss function to reduce the influence of occluded pixels and inter-frame motion which results in out-of-bounds regions on depth information prediction during pixel projection. By improving the depth estimation network, pose estimation network, and loss function, the accuracy of the unsupervised monocular depth estimation model for depth information is improved, and the robustness and generalization performance of the model is enhanced.

The main contributions of our works are as follows:

- We propose an unsupervised monocular depth estimation method improved for ill-posed regions by training a depth estimation network and a pose estimation network in cascade with loss functions.
- We improve the unsupervised depth estimation network by using asymmetric convolution, multiscale perceptual field structure, SE structure and minimum reprojection error in ill-posed regions, such as weak texture scenes, pixel projection occlusion, lack of detailed information in depth images, and so on.
- Our approach demonstrate state-of-the-art performance at KITTI monocular datasets.

2 Approach

At present, the unsupervised monocular depth estimation model takes video sequences as input and constructs an unsupervised learning framework for monocular depth and camera pose estimation based on unstructured video sequences. Specifically, an end-to-end learning method is used to jointly train a depth estimation network and a pose estimation network in an encoder-decoder manner, so as to obtain the depth information in a single frame of a video sequence in an unsupervised manner [11].

However, current unsupervised monocular depth estimation algorithms still have limitations when dealing with ill-posed regions, such as weak texture scenes, occlusion of pixel projection, detail information lack of depth images, and redundant errors of continuous image frames for pose information.

In order to further improve the unsupervised monocular depth estimation model and cope with the above complex scenes, this paper improves the unsupervised monocular depth estimation model, which consists of depth estimation network, pose estimation network, and the loss function. We predict the depth information and pose information of 2D images by cascading the depth estimation network and pose estimation network, then we take the pixel error between the reconstructed image and the input image as the supervised signal of the whole network to achieve the depth estimation of unsupervised monocular estimated images. Firstly, for the depth estimation network, inspired by Ding [11], the AC (Asymmetric Convolution) is designed to extract the features of the input image from vertical, horizontal, and overall directions, so as to alleviate the influence of weak texture scenes. Through RFB (Receptive Field Block) which is a multi-scale receptive field structure [12], the ability to obtain all and local information is enhanced in the receptive field area of different scales of the network. Secondly, for the pose estimation network, SE(Squeeze-and-Excitation) structure [13] is introduced to reduce the error region of pose estimation. Finally, for the loss function, the concept of minimizing reprojection error is introduced to reduce the impact of pixel projection occlusion in depth information estimation.

The overall structure of improved unsupervised monocular depth estimation network is shown in Figure 1. Firstly, multi-scale feature maps which is equivalent to 1/2, 1/4, 1/8, 1/16 resolution of the input image frame are generated in the improved depth estimation network, and then these features are mapped to the depth decoder with parameter sharing, and the estimated depth is restored to the same size as the resolution of the input image frame through the upsampling structure. Secondly, for the improved pose estimation network, the relative pose of 6 degrees of freedom which includes displacement with 3 degrees of freedom and spatial rotation with 3 degrees of freedom is generated by the pose estimation network. Finally, the depth information and pose information obtained by the improved depth estimation network and pose estimation network are jointly trained using the loss function.

2.1 The depth estimation network optimization

At present, most unsupervised monocular depth estimation algorithms cannot effectively deal with weak texture scenes and miss detailed information of the predicted depth image. In order to solve this problem, asymmetric convolution and multi-scale receptive field RFB are used in the depth estimation network to enhance the recognition of weak texture scenes and strengthen the acquisition of detailed information. The depth estimation network is improved accordingly.

2.1.1 Improved ACResNet50 depth estimation network

Weak texture regions are not distinct and significant features, which are prone to semantic ambiguity and lead to wrong depth

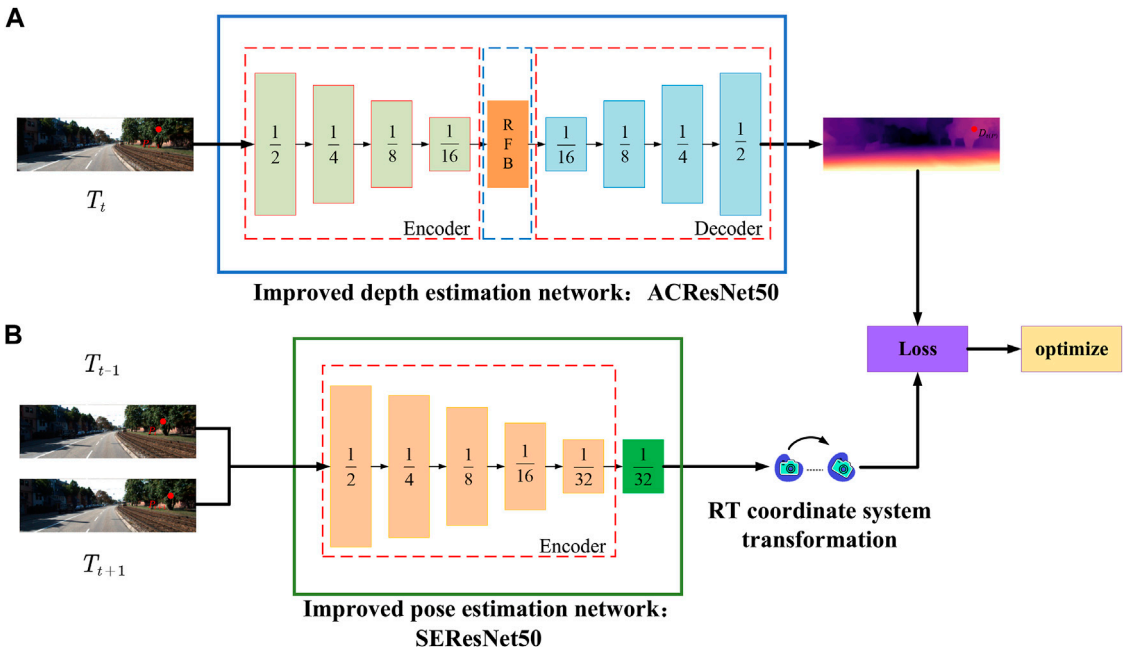


FIGURE 1
Structure diagram of unsupervised monocular depth estimation model: **(A)** Improved depth estimation network ACResNet50; **(B)** Improved pose estimation network SEResNet50.

estimation, so we deal with this problem in this paper. Our improved ACResNet50 depth estimation network is shown in Figure 2.

Firstly, in order to effectively mitigate the impact of weak texture scenes on the accuracy of depth estimation information, the traditional convolution method is replaced by AC, and each traditional convolution of ResNet50 network is replaced to strengthen its feature extraction ability, and the ACResNet50 network structure is formed. Secondly, in order to solve the problem of missing details, the RFB structure is connected to the last structure of ACResNet50. Based on the convolution of different sizes, the dilated convolution is added and its expansion rate is adjusted to ensure the network receptive field, so as to achieve the acquisition of high-resolution features. The fusion of global feature information and local feature information is strengthened. Finally, the deconvolution structure is used to restore the size of the output feature map to the size of the original feature image, and the prediction function of the entire network depth information is realized.

2.1.2 Asymmetric convolution

At present, the unsupervised monocular depth estimation network performs poorly on weak texture scenes. Most unsupervised monocular depth estimation networks use the ResNet50 network as the feature extraction backbone network in the encoding process of the image and extract the feature information of the image by feature superposition and refinement. Although the residual structure of ResNet50 can extract the feature information of the image to a certain extent, it is far from sufficient for the task of unsupervised monocular depth estimation that requires more accurate depth information. At the same time, the continuous superposition of the ResNet50 network and the deepening of the number of network layers will also lead to many problems, such as too many network parameters, difficult training, and the degradation of the whole network.

In order to obtain more feature information of the input image and alleviate the influence of weak texture scenes on unsupervised monocular depth estimation tasks, inspired by ACNet research, the traditional convolution method is improved, and we propose a novel depth estimation network based on ACNet. The feature extraction of the input image is carried out from the vertical, horizontal, and overall directions, which strengthens the feature information extraction ability of the feature extraction network and alleviates the influence of weak texture scenes on the depth information.

Figure 3 is the operation process of asymmetric convolution. The ACNet network in the Figure 3 can be divided into two stages, training and test reasoning, with Figure 3A indicating the training stage and Figure 3B indicating the test reasoning stage. Firstly, we set up three parallel convolution kernels with sizes 1×3, 3×1, and 3×3 respectively, 1×3 and 3×1 convolution kernels facilitate the extraction of edge information of weak texture regions and other regions to identify weak texture regions with other regions, and then joint 3×3 convolution to extract contextual features of weak texture regions to improve the accuracy of weak texture depth estimation. Secondly, the input image is processed by these three parallel convolution kernels respectively, so that the extracted feature information has the characteristics of horizontal, vertical, and overall directions, then the three kinds of feature information can be stacked and output. Finally, the traditional convolutions in the network are replaced with non-traditional convolutions to form the improved feature map extraction network on ResNet50.

2.1.3 Multiscale receptive fields

The lack of details in depth maps has always been a difficulty for unsupervised monocular depth estimation. The reason is that in the

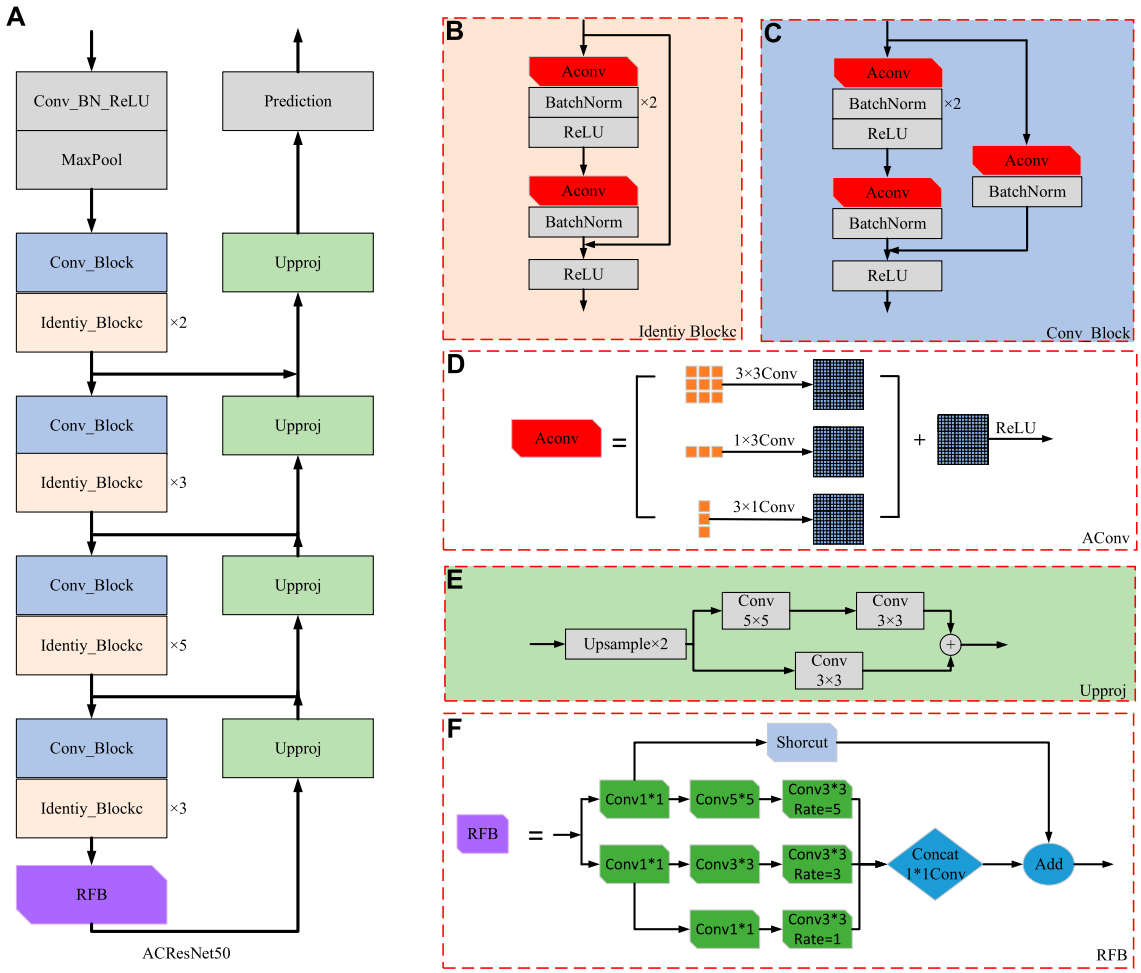


FIGURE 2
Depth estimation network model: **(A)** ACResNet50 structure; **(B)** Identiy_Block structure; **(C)** Conv_Block structure; **(D)** AConv structure; **(E)** Upproj structure; **(F)** RFB structure.

theory of deep convolutional neural networks, the perceptual field of the network gradually increases with the number of layers of the network, Zhou [14] founds that the network's ability of detail acquisition in the receptive field is reduced in deeper networks, leading to poor network learning. Moreover, in the traditional convolution process, convolution is used to continuously stack down sampling to extract abstract information, but continuous downsampling will lead to the loss of image details and local information. Zhao [15] points out that the fusion of global and different scale context information in semantic segmentation is beneficial to alleviate the loss of detail information and preserve the spatial structure of the image. Therefore, RFB is adopted to solve this problem in that the receptive field decreases in the unsupervised monocular depth estimation model, which leads to unsatisfactory context information fusion and missing details of the estimated depth map.

The RFB can achieve the acquisition of high-resolution features without repeated down-sampling and enhance the ability of network feature extraction and fusion [12]. At the same time, different receptive fields are obtained by adjusting different expansion rates of dilated convolution, so as to enhance the variability of network receptive field region size. By stacking in this way, the ability of interfusion between feature information at different scales of the network is enhanced, and the acquisition of full and local detail information is strengthened. The multiscale receptive field RFB structure is shown in Figure 4.

In this paper, a multi-scale receptive field RFB structure is added after the last convolutional block of the ACResNet50 feature extraction network. Firstly, in the multi-branch convolution layer, convolution kernels of 1×1, 3×3 and 5×5 sizes are used to ensure the performance of the network to deal with scale changes and improve the multi-scale feature extraction ability of the model. Secondly, on the dilated convolution, in order to ensure the consistency of the scale of the multi-branch convolution layer and the expansion rate of the dilated convolution, by connecting cavity convolution with expansion rates of 1, 3 and 5, respectively to convolution of different scales, we enhance the receptive field of the network and improve the acquisition ability of high-resolution feature maps and context information. Finally, the image feature information of different scales is fused by stacking the features to generate a receptive field spatial array as the input of deconvolution through 1×1 convolution.

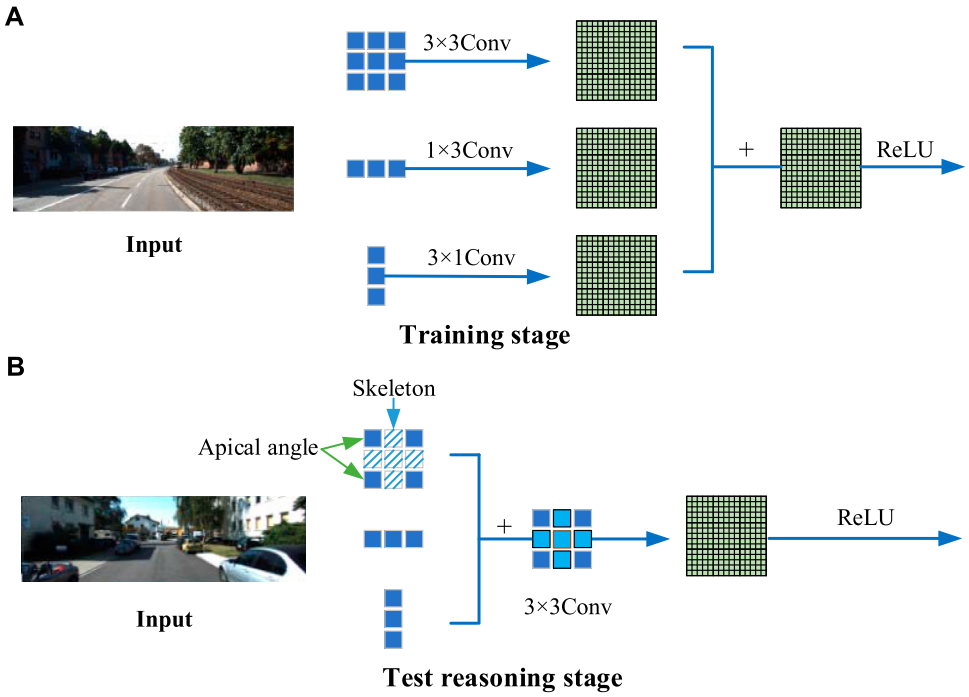


FIGURE 3
Asymmetric convolution structure:**(A)** The ACNet structure in the training stage; **(B)** The network structure in the test reasoning stage.

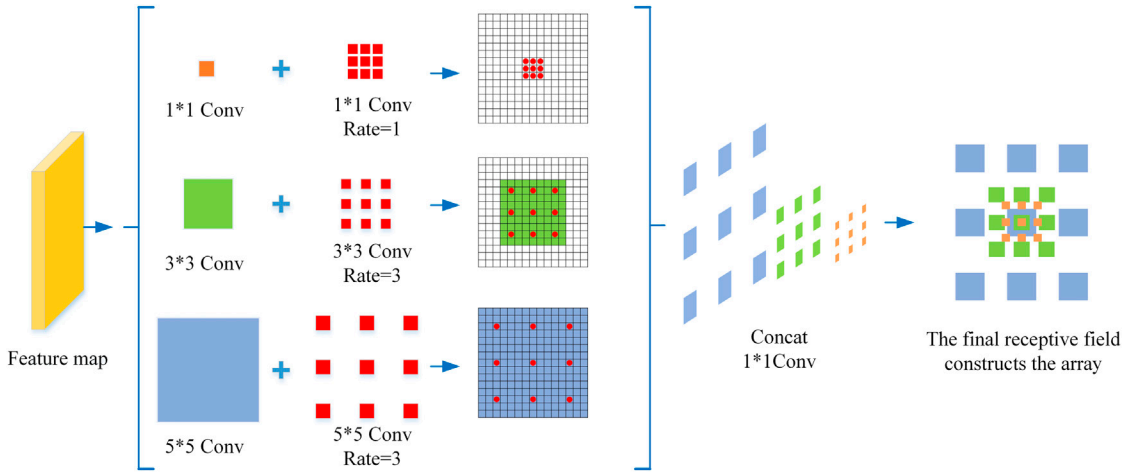


FIGURE 4
Multi-scale receptive field RFB structure.

2.2 Pose estimation network based on SEResNet50

The pose estimation network is crucial for accurately predicting depth information. However, in the design process of the pose estimation network, most unsupervised monocular depth estimation models directly use the pose information of consecutive image frames for model prediction, ignoring the redundant error of pose information, which leads to the reduction of the accuracy of the model prediction depth information.

In order to reduce the large redundant errors in pose estimation, we design the SE attention mechanism structure based on ResNet50 in the pose estimation network, which can focus on the important pose information of the image frame, suppress the unimportant pose information of the image frame, and reduce the large error redundancy. The improved pose estimation network is shown in Figure 5.

2.2.1 SE attention mechanism

For the pose estimation network, its task is to accurately predict the camera motion trajectory between adjacent frames in the video

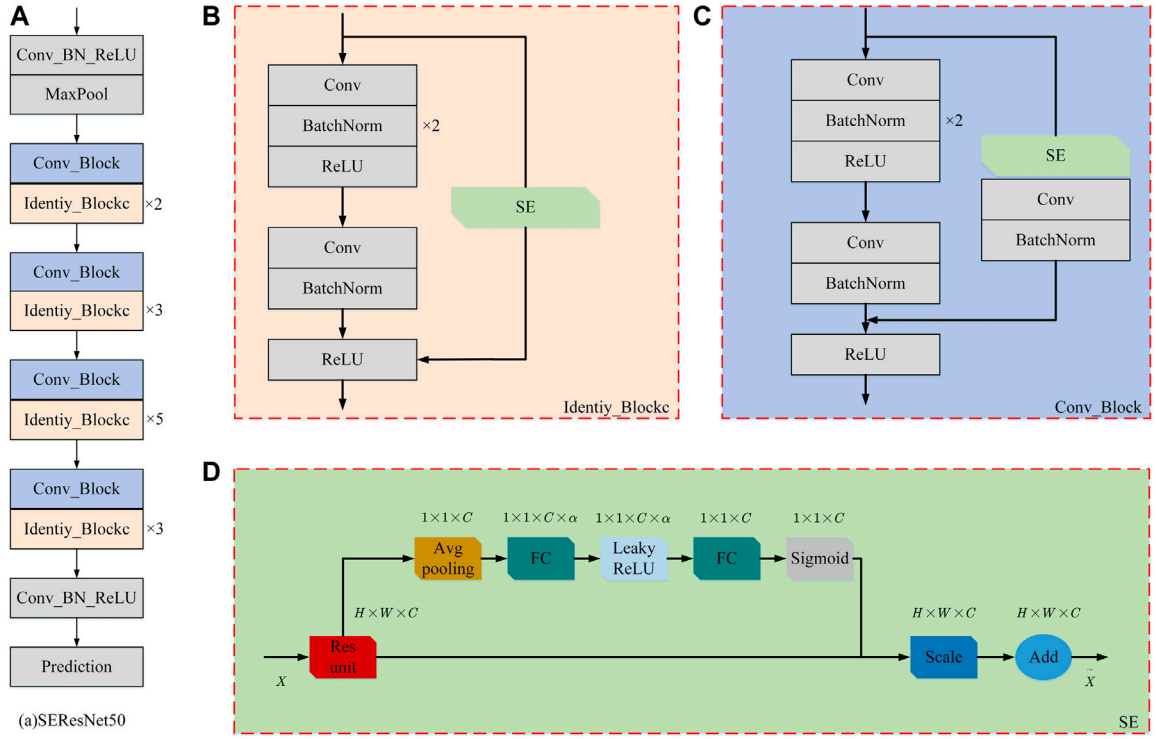


FIGURE 5
Improvement of pose estimation network model: **(A)** SEResNet50 structure; **(B)** Identiy_Block structure; **(C)** Conv_Block structure; **(D)** SE structure.

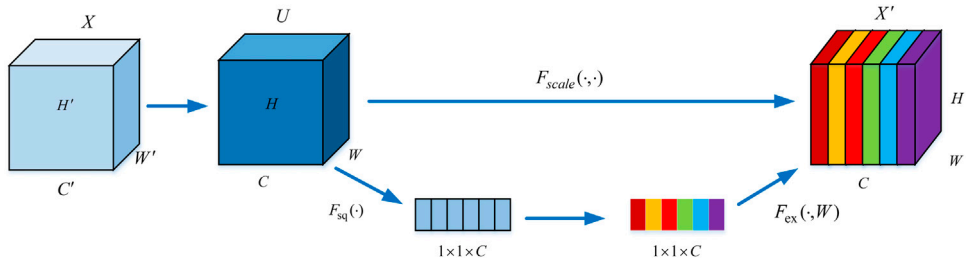


FIGURE 6
SE channel attention mechanism.

sequence, so as to obtain the rotation matrix and translation matrix. Then, the image is reconstructed by combining the internal parameter matrix of the camera and the depth information predicted by the depth estimation network. However, in the pose estimation network, the camera pose motion estimation in the image between two highly adjacent frames is highly approximate. If the network trains all the pose information of the video sequence frames and predicts the camera pose, it will not only increase the amount of information processed by the network but also lead to an increase in redundancy error in the pose estimation.

SE structure focuses on exploring the relationship between different channels in feature information, and this exploration method has a good performance in balancing the importance of feature channels and learning global feature information [13]. In the pose estimation network, the SE structure can be used to pay more attention to the important pose information in the continuous image frames of the video sequence, suppress the unimportant pose information, and effectively enhance the network's prediction of the camera pose motion trajectory between image frames, and improve the ability of pose estimation.

Figure 6 shows the attention mechanism structure of SE channel. Firstly, given a feature input X, its height, width, number of channels, and the dimension are H', W', C', and $H' \times W' \times C'$, respectively. After a series of transformations such as convolution, a feature U of size $H \times W \times C$ is obtained. Secondly, the squeeze operation $F_{sq}(\cdot)$ is carried out, so that the feature U is squeezed along the spatial dimension. Further, each two-dimensional characteristic channel is turned into a real number, and a feature with the same dimension and channel number is output, whose size is $1 \times 1 \times C$. Thirdly, the excitation operation $F_{ex}(\cdot, W)$ is used to generate a weight for each feature channel, where W represents the correlation between feature channels. Finally, by doing the scale operation $F_{scale}(\cdot,\cdot)$, the weight

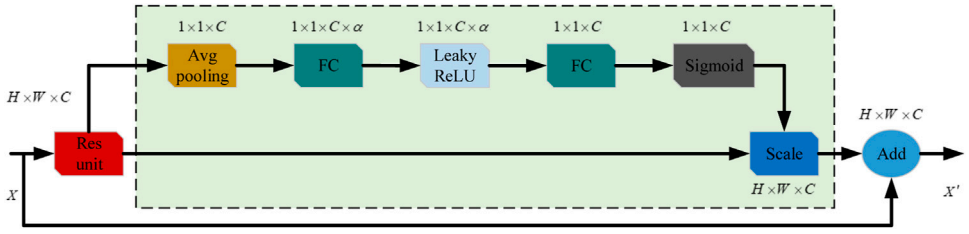


FIGURE 7
SE+ Residual blocks structure.

output by the excitation operation is weighted to the feature U of the previous layer channel by channel through multiplication, and the feature X' with attention mechanism is obtained with size $H \times W \times C$. Through the above operations, we retain important feature information and strengthen learning ability of global feature information.

2.2.2 Residual network with attention mechanism

There are multiple block structures in the original ResNet50 feature extraction network, and each block realizes the extraction of image features by stacking each other. The original ResNet50 backbone feature extraction network does not consider the relationship between different channels in the feature information. Such a way will lead to the lack of ability to distinguish the main and secondary channel feature information, resulting in a weak performance in global feature information extraction ability.

However, the SE channel attention mechanism makes full use of the weights of different feature channel information importance to enhance the information acquisition of important feature channel. In this paper, the attention mechanism is introduced into the backbone feature extraction network to strengthen the extraction performance of global feature information. Its improved residual network with attention mechanism is shown in Figure 7.

This paper designs the channel attention mechanism SE in each block structure of the ResNet50 feature extraction network. Firstly, the feature map of height H, width W, number of channels C and size $H \times W \times C$ by ResNet50 is output a feature map of size $1 \times 1 \times C$ by global average pooling. Secondly, the feature map of $1 \times 1 \times C$ is input to the first fully connected layer with ReLU as the activation function and the output is $1 \times 1 \times C \times \alpha$, where the number of neurons is $C \times \alpha$ and the scaling parameter is α , which aims to reduce the channel reduction calculation. It is input to the second fully connected layer with sigmoid as the activation function, whose output is $1 \times 1 \times C$ and the number of neurons is C to complete the acquisition of the weight of the attention mechanism of different channel feature information. Then, the obtained weights are applied to the $H \times W \times C$ feature information of ResNet50 output through the multiplication operation to obtain the feature channels with weights. Finally, the feature output of the previous layer and the weighted feature channel are superimposed to obtain the final feature output. An improved SEResNet50 residual block with attention mechanism is formed. This structure can effectively enhance the performance of the network in extracting feature information of important adjacent frame image pose changes and reduce the redundancy error of the pose estimation network.

2.3 Design of the loss function

In the design of the loss function, since the whole unsupervised monocular depth estimation network consists of two parts: the depth estimation network and the pose estimation network, which are used together to predict the depth of a pixel. Therefore, the constraint term of the loss function is derived from the pixel difference between the reconstructed image and the input image after information predicted by the depth estimation network and the pose estimation network. In the inference of the loss function, let the three adjacent frames of images at time t be I_t, I_{t-1}, and I_{t+1}.We call I_t the target image and the other two I_{t-1} and I_{t+1} the source images. Firstly, the depth $\hat{D}_t(p_t)$ of each pixel p_t in the target view I_t is obtained through the depth estimation network, and then (I_t, I_{t-1}) and (I_t, I_{t+1}) are fed into the pose estimation network as a group to obtain the camera motion $\hat{T}_{t\to t-1}$ and $\hat{T}_{t\to t+1}$ between neighboring pixels respectively. In this way, the depth information and pose information of the color image are obtained.

In the process of image reconstruction, each pixel p_t in the target view I_t is projected onto the source image $I_{s_\wedge} \in (I_{t+1}, I_{t-1})$ at pixel s according to the predicted depth information $\hat{D}_t(p_t)$ and camera pose $\hat{T}_{t\to t-1}, \hat{T}_{t\to t+1}$. Bilinear interpolation is then used to obtain p_t which is the value of the distorted image. The differentiable image warping process is shown in Figure 8.

For the flush coordinate p_t of a pixel in the target frame, then the projection coordinate of p_t corresponding to the p_s of the source frame can be obtained as follows:

$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t, \tag{1}$$

where $\hat{T}_{t\to s}$ is the camera motion pose from frame t to s, $\hat{D}_t(p_t)$ is the depth value of pixel p_t in frame t, and K is the camera internal reference matrix.

In this case, let the target image I_t of the reconstructed frame, the source image I_s as the frame used to reconstruct I_t, and the reconstructed image $\hat{I}_s$. Let $< I_1, \ldots, I_N >$ be a training image sequence, where one of the frames is denoted as the target image I_t. I_s is the source image sequence denoted as $I_s (1 \le s \le N, s \ne t)$. $||$ measures the absolute error. Then the loss function L_1 is expressed as follows:

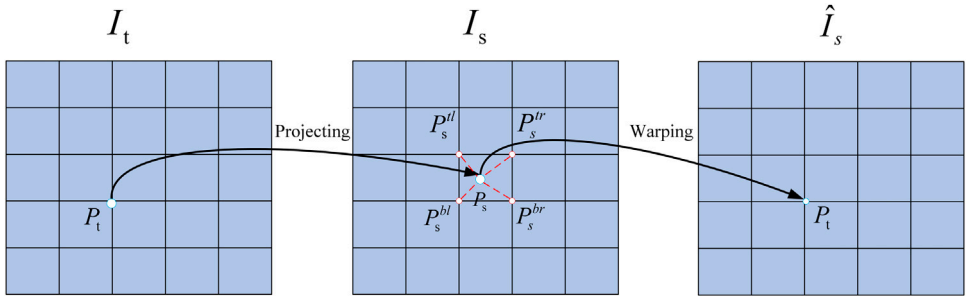


FIGURE 8
Differentiable image warping process.

$$L_1 = \sum_s \sum_p \left| I_t(p) - \hat{I}_s(p) \right|. \tag{2}$$

Since the premise assumptions of invisible change and static scene need to be satisfied in network construction, if one of the assumptions is not met, the gradient will be destroyed and the inhibition of training will occur. In response to these factors, in order to improve the robustness of the network, the output confidence weight $\hat{E}_S(p)$ for each target source pair is given during the cascaded training of the depth estimation network and the pose estimation network. After weighting the loss function (2), the loss function L_2 is expressed as:

$$L_2 = \sum_s \sum_p \hat{E}_s(p) \left| I_t(p) - \hat{I}_s(p) \right|. \tag{3}$$

In the original algorithm, the ill-posed region is solved by adding a smoothing constraint when obtaining the depth map, and the depth of each pixel is solved by global optimization. However, this method is to average the reprojection error of multi-source images, which may lead to problems of pixels which are visible in the target image and invisible in the source image. If the network predicts the correct depth of a pixel, then the corresponding color in the blocked source image has a high probability of mismatch with the target, resulting in a high photometric error. There are two reasons for this problem. One is pixels which are on the edge of the image and are out of view due to motion between frames. The other is the occluded pixels.

In this paper, we use the concept of minimum reprojection error to deal with the problem of out-of-bounds caused by occluded pixels and inter-frame motion. At each pixel, the photometric error of all source images is no longer averaged, but simply the minimum value is used, which can effectively alleviate the pixels that are visible in the target image and invisible in the source image in the process of pixel projection, and solve the occlusion problem caused by pixel projection. Therefore, the calculation process of the minimum reprojection loss function L_p is as follows:

$$L_p = \sum_{t'} pe(I_t, I_{t' \to t}), \tag{4}$$

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - SSIM(I_a, I_b)) + (1 - \alpha)\|I_a - I_b\|_1. \tag{5}$$

Among them, SSIM (Structural Similarity Index Measurement) is the structural similarity index, I_a and I_b are the adjacent frame images, t represents the time of each frame image. As known relative pose at time t', the source image $I_{t'}$ is the second frame in the stereo pair to I_t, and α is set to .85 to make its edge perception smooth.

Finally, the minimum reprojection error constraint is introduced into the overall loss function to reduce the impact of pixel occlusion on the model during pixel projection and ensure the accuracy of the model in predicting depth information. The final loss function of the model L_{final} is as follows:

$$L_{final} = \sum_l L_1^l + \lambda_p L_p^l + \lambda_e \sum_s L_{\text{reg}}\left(\hat{E}_s^l\right). \tag{6}$$

λ_p and λ_e are the weight value of minimizes the reprojection error and the weight value normalized by the target source on the output confidence. We empirically take the values are .65 and .35, respectively. L_{reg} denotes the regularization term [16], and l represents different image scales, respectively.

3 Experiments

In our experiment, video images of real scenes are utilized as training data set and test data set, such as urban areas and highways in KITTI data set. In order to ensure the consistency of the experiment, the image resolution is uniformly cropped to a size of 640 × 192. The common methods of data enhancement such as rotation and flip are also used to expand the data. The SGD (Stochastic Gradient Descent) algorithm is used to optimize the model parameters. The training iteration epochs of the whole network is set to 200. The initial learning rate is set to .001 and dynamic attenuation is adopted. Image acceleration is CUDA11.2.0/CUDNN8.2.1.

3.1 The ablation experiment

To verify the reliability of the proposed scheme, we validated the proposed scheme on the KITTI dataset and performed ablation experiments and compared the proposed method in this paper with the scheme of Zhou [10], and the experimental scheme and results are shown in Table 1.

3.2 The depth estimation network

3.2.1 Verification of asymmetric convolution structure

In order to verify AC, we conduct comparative experiments between ACResNet50 in this paper and Zhou's method.

From the quantitative and qualitative analysis of the relevant evaluation indicators in Table 1 and Figure 9, our method works

TABLE 1 Ablation experimental design protocol and comparison of experimental results.

Category of schemes					Error metric			Accuracy metric		
Zhou [10]	ACResNet50	RFB	SEResNet 50	L_p	Abs rel	Rmse	Rmse log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
√	×	×	×	×	0.183	6.709	0.27	0.734	0.902	0.959
√	√	×	×	×	0.169	6.391	0.262	0.740	0.91	0.963
√	√	√	×	×	0.164	6.249	0.258	0.758	0.915	0.965
√	√	√	√	×	0.162	6.211	0.246	0.773	0.918	0.968
√	√	√	√	√	0.161	6.032	0.235	0.781	0.922	0.970

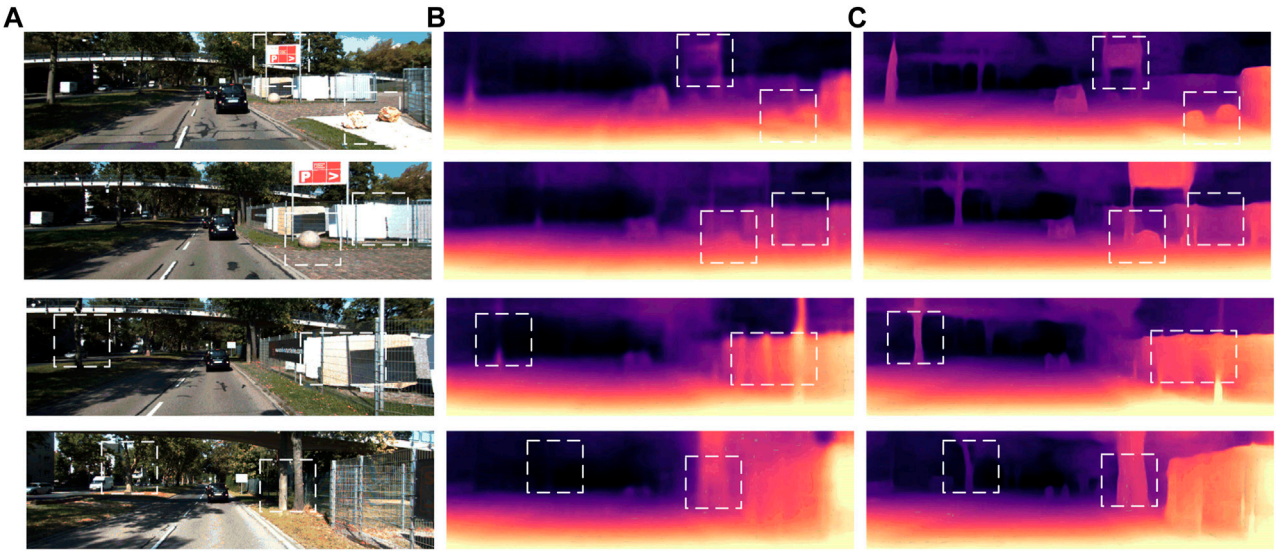


FIGURE 9
Comparative experimental results of asymmetric convolution module visualization on KITTI dataset: **(A)** represents the color image input by the model; **(B)** represents the depth map result predicted by Zhou et al.; **(C)** The ACResNet50 prediction of asymmetric convolution to the depth map results.

better than Zhou's method for weak texture scenes in the ill-posed regions of the billboard in row 2 and the columnar objects in rows 2 to 3. Only using the asymmetric convolution structure designed to replace the traditional convolution structure has a certain improvement in the accuracy of the estimated depth value of the network. The experimental results show that the improved asymmetric convolution can effectively enhance the ability of the network to obtain feature information for the color two-dimensional image, strengthen the feature extraction of the input image, and make the unsupervised monocular depth estimation network output depth images with rich textures and clear edges.

3.2.2 Validation of ACResNet50+ RFB structure

In order to verify the RFB structure, the ACResNet50 + RFB is compared with Zhou [10].

The relevant quantity and quality evaluation metrics are analyzed in Table 1 and Figure 10. In this paper, RFB is introduced into the last module of the ACResNet50 network, so that the model can obtain the context information of image features at different scales. The obtained feature information is more continuous and the detail information is more complete, which ensures the continuity and integrity of the spatial structure of the output depth image of the network. In Figure 10, our method is able to retain more detailed information of vehicle contours, which is significantly better than Zhou's method. Experimental results show that the proposed multi-scale receptive field enhanced RFB structure outperforms Zhou's algorithm in depth map detail information and spatial structure presentation. It can effectively avoid the lack of details in the unsupervised monocular image depth estimation task, strengthen the control of the model for detailed information. At the same time, it can further obtain multi-scale information and rich context information in two-dimensional color images, and improve the overall prediction accuracy and generalization performance of the model. The results show that the method can effectively alleviate the redundancy error problem of detail information in the ill-posed regions.

3.3 Pose estimation network

In order to verify the actual effect of the pose estimation network SEResNet50 embedded with the attention mechanism designed in this paper, the method in this paper is compared with Zhou [10].

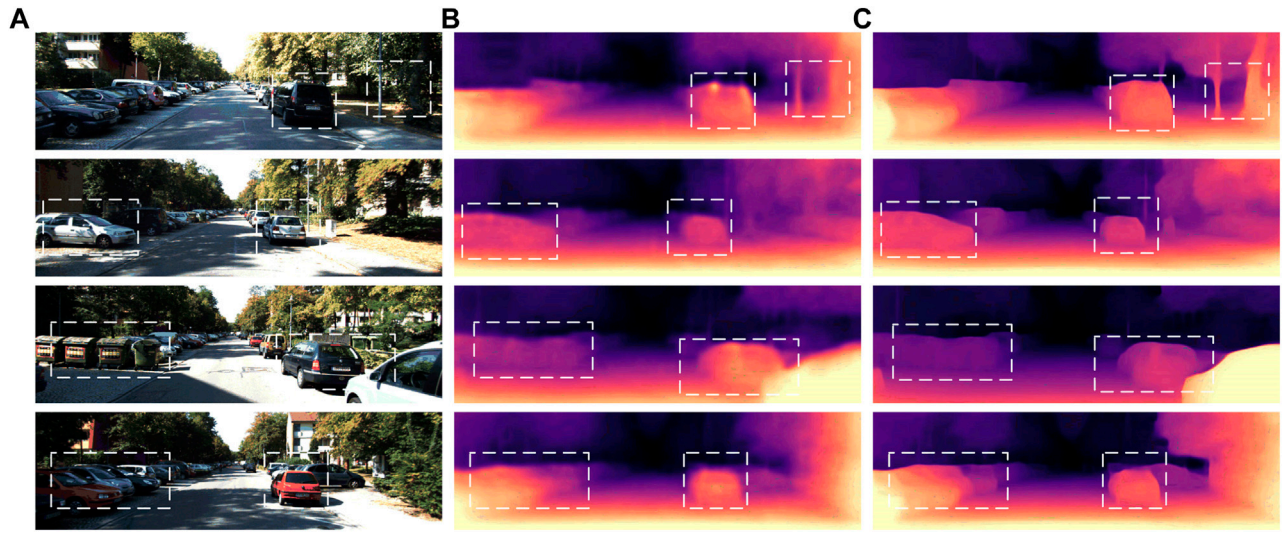


FIGURE 10
Comparative experimental results of multi-scale receptive field RFB visualization on KITTI dataset: **(A)** The color images input by the model; **(B)** The depth map results predicted by Zhou et al.; **(C)** The depth map results predicted by ACResNet50+ RFB structure.

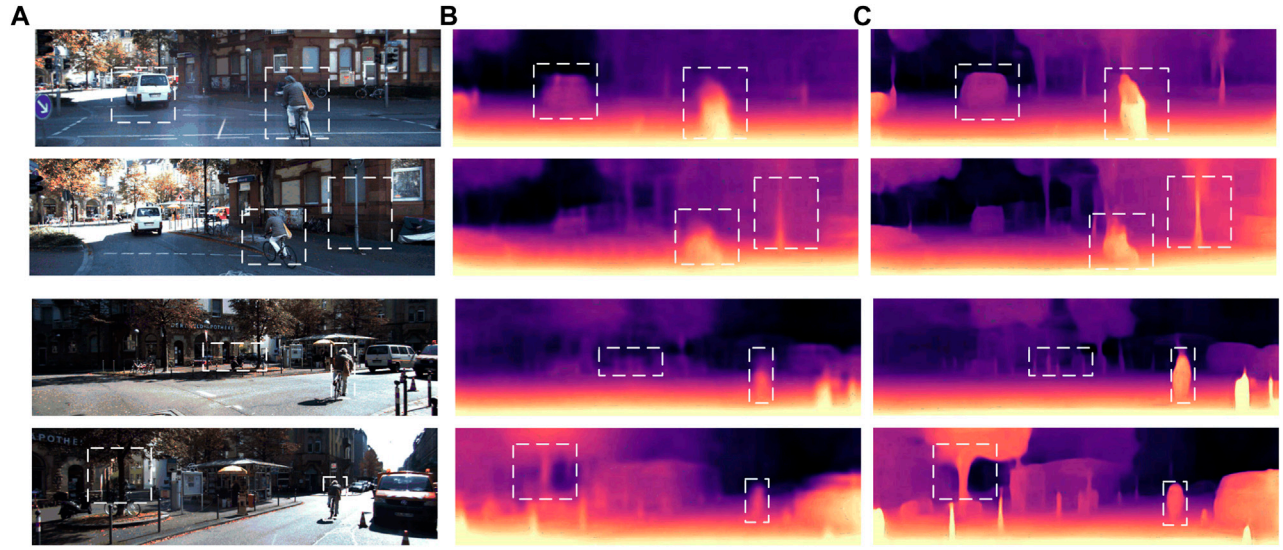


FIGURE 11
Comparative experimental results of attention mechanism SE module visualization on KITTI dataset: **(A)** The color image input by the model; **(B)** The depth map results predicted by Zhou et al.; **(C)** The depth map predicted by SEResNet50 embedded attention mechanism in the pose estimation network.

The relevant evaluation indicators in Table 1 and Figure 11 are analyzed quantitatively and qualitatively. In this paper, the attention mechanism SE structure is designed to reduce the redundant error caused by using the pose information of consecutive frames to predict the pose information of the next frame in the pose estimation process. The attention mechanism SE can pay attention to the important information in a single frame and suppress the unimportant information, so as to effectively reduce the redundant error generation and improve the overall prediction accuracy of the model. From Figure 11, we can find that our method works well when targeting the projected occlusion region of bicycle pedestrians and car outline. The experimental results show that the attention mechanism SE structure designed in this paper can reduce the redundant error of camera pose estimation in the pose estimation network. In terms of the accuracy of predicting the depth value, the three indicators have a corresponding improvement, where $\delta < 1.25$, it is an obvious improvement over Zhou [10], and the output depth map is of high quality. It shows that the pose estimation network designed in this paper can effectively estimate the motion pose of the camera accurately, and it is a good contribution to the whole unsupervised monocular depth estimation network to predict depth information.

At the same time, in order to further verify the absolute trajectory error estimated as the pose information, the prediction results are

TABLE 2 Absolute trajectory error for validating positional estimation on KITTI test set.

Methods	Seq.9	Seq.10
ORB-SLAM (full) [17]	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short) [17]	0.064 ± 0.141	0.064 ± 0.130
Zhou [10]	0.021 ± 0.017	0.020 ± 0.015
Our method	0.019 ± 0.015	0.018 ± 0.016

tested through the pose estimation test data Seq.9 and Seq.10 provided by the KITTI dataset official website, as shown in Table 2.

As can be seen from Table 2, after designing the attention mechanism in the pose estimation network, the error of pose estimation on the KITTI test set is smaller than that of ORB-SLAM (short) and Zhou's method, but larger than that of ORB-SLAM (full). Therefore, the attention mechanism used in the pose estimation network can effectively reduce the redundant error caused by the superposition of consecutive multi-frame image information and improve the robustness of the model.

3.4 Minimum reprojection error loss function

In order to verify the experimental effect of introducing the minimum reprojection error loss function. The model introduced with the minimum reprojection error loss function designed in this paper is compared with the method of Zhou [10].

The relevant evaluation indicators in Table 1 and Figure 12 are analyzed quantitatively and qualitatively. In this paper, a constraint term of minimum reprojection error is added to the loss function, which is beneficial for the prediction of depth information, and can effectively improve the occlusion problem in the projection process of adjacent pixels.

Experimental results show that after using the minimum reprojection error as a constraint term, each error index is reduced accordingly. It improves the problem of occlusion during the projection of adjacent pixels and enhances the prediction accuracy of depth information of the model. At the same time, the robustness and generalization performance of the model are improved.

3.5 KITTI contrast experiment

At the same time, in order to verify the effectiveness and generalization of the proposed method, we make qualitative and quantitative comparison analysis with the research algorithms in related fields. In order to verify the effectiveness of the method in this paper, the comparative experiments are based on the KITTI dataset, verify the generalization of the method in this paper, the cityscapes dataset is used, but the error of the model increases slightly when dealing with data sets other than KITTI.

In Table 3, k is the KITTI dataset, CS is the Cityscapes dataset, and supervision (Y, N) indicates whether it is an unsupervised and supervised monocular depth estimation task. The relevant evaluation indicators in Table 3 and Figure 13 are analyzed quantitatively and qualitatively. The algorithm designed in this paper is .022, .677, and .035 lower than Zhou in AbsRel (Absolute Relative error), RMS (Root Mean Square error), and LogRMS (Log Root Mean Square error), respectively. In the three depths value accuracy evaluation indicators of $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$, it is .047, .020, and .013 higher, respectively. The accuracy of predicting depth information from monocular color image is also better than that of the algorithm proposed by Zhou [10].

The method designed in this paper has good performance in various evaluation indicators compared with the previous research work. Among them, compared with the supervised method of Eigen [18], Liu [19] and Cao [22], the accuracy of the predicted depth value is greatly improved. Compared with the unsupervised monocular depth estimation proposed by Zhou [10], the three indexes in this paper are increased by .047, .020, .013 respectively, and the error index is reduced accordingly. Compared with the recent work of Yang [21], AdaDepth [23], S2R-DepthNet [24], etc. which studied the unsupervised monocular depth estimation task, the proposed method performs better in all indicators. At the same time, from the depth images predicted by each algorithm in Figure 13, the proposed algorithm has good performance in the texture information, detail information, and spatial structure of the output depth map.

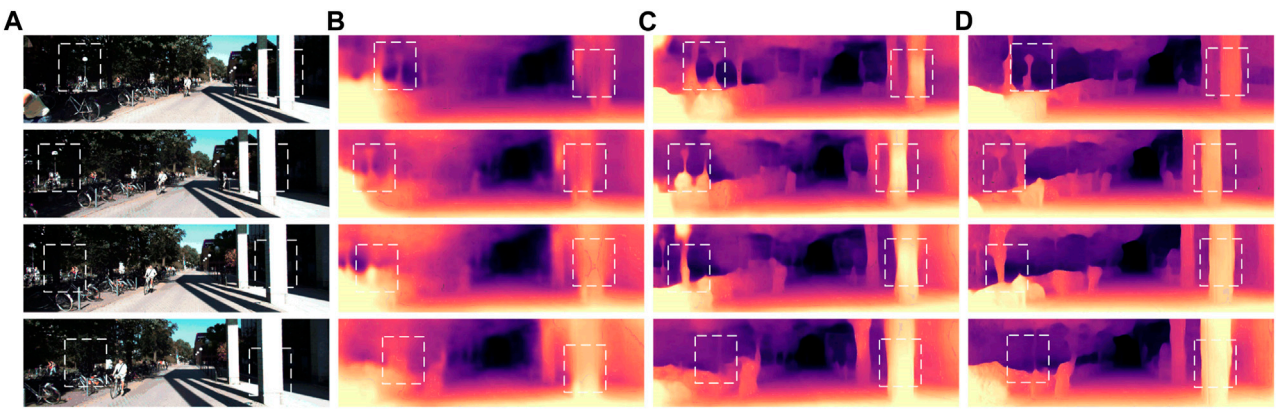


FIGURE 12
Comparative experimental results of the per-pixel minimum reprojection error visualized in the KITTI dataset: **(A)** The color image input by the model; **(B)** The depth map result predicted by Zhou et al.; **(C)** The depth map predicted by the whole network structure after improvement; **(D)** The depth map predicted by the whole model after using the minimum reprojection error loss function.

TABLE 3 Comparison of experimental results with other related research algorithms.

Methods	Supervised	Data	Error			Accuracy, δ		
			AbsRel	RMS	LogRMS	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen [18]	Y	K	0.214	6.307	0.292	0.673	0.884	0.957
Liu [19]	Y	K	0.202	6.471	0.275	0.678	0.895	0.965
Zhou [10]	N	K	0.183	6.709	0.27	0.734	0.902	0.959
UnDeepVO [20]	N	K	0.183	6.57	0.268	—	—	—
Yang [21]	N	K	0.182	6.501	0.267	0.725	0.906	0.963
Cao [22]	Y	K	0.180	6.311	—	0.771	0.917	0.966
AdaDepth [23]	N	K	0.167	5.578	0.237	0.771	0.922	0.971
S2R-DepthNet [24]	N	K	0.165	5.695	0.236	0.781	0.931	0.972
Geonet [25]	N	K	0.164	6.09	0.247	0.765	0.919	0.968
Mahjourian [26]	N	K	0.163	6.22	0.25	0.762	0.916	0.966
LEGO [27]	N	K	0.162	6.276	0.252	—	—	—
Our method	N	K	0.161	6.032	0.235	0.781	0.922	0.972
Our method	N	CS	0.174	6.322	0.259	0.748	0.911	0.964
Our method	N	K + CS	0.168	6.282	0.26	0.731	0.908	0.963

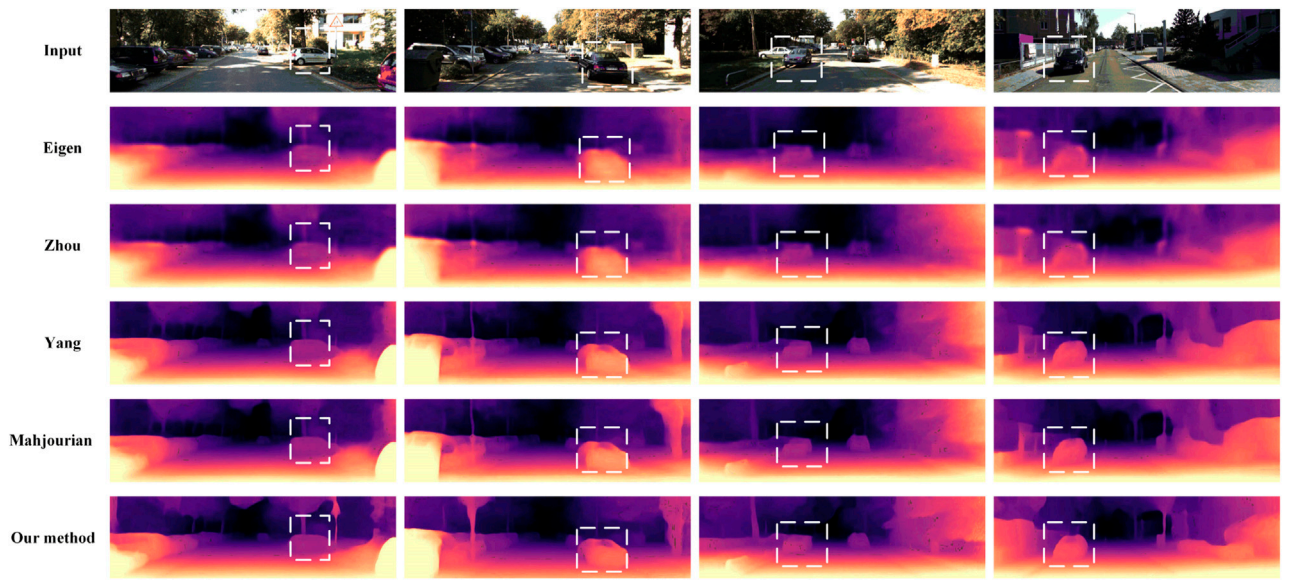


FIGURE 13
Visual comparison experimental results with other related research algorithms on the KITTI dataset: the first row is the input two-dimensional color image; rows 2 to 6 show the depth maps predicted by Eigen, Zhou, Yang, Mahjourian et al. and our method, respectively.

The experimental results show that the improved unsupervised monocular depth estimation algorithm designed in this paper can effectively alleviate the impact of weak texture scenes on the model, solve the lack of detail of the input image, reduce the redundant error of pose information, reduce the occlusion problem in the process of pixel projection, and ensure the prediction accuracy of the unsupervised monocular depth estimation model. From the analysis of the above indicators, the unsupervised monocular depth estimation network has a certain competitive advantage in depth prediction, and can accurately estimate the depth information of images or video frames.

4 Conclusion

Currently, supervised monocular image depth estimation tasks require a large amount of real depth data for training, which greatly

increase the development cost of the model and the difficulty of landing the model. The improved unsupervised monocular depth image estimation task designed in this paper only uses continuous video sequences to complete the depth prediction of each pixel of a single image, which greatly reduces the model development cost and accelerates the model implementation process. It can effectively improve the influence of weak texture scene on depth prediction, reduce the lack of details of the model predicted depth image, and reduce the occlusion problem of the model due to the pixel projection process. Through the improvement of this paper, the prediction accuracy of the unsupervised monocular image depth estimation model on depth information is strengthened, which makes the depth image predicted by the model richer in texture information, clearer in detail information, and more continuous in spatial structure, thus enhancing the structure of the predicted depth image and improving the resolution of the output image. The robustness and generalization performance of the unsupervised monocular depth estimation model are improved.

Although our approach does not require labeling of real depth images as supervised methods do, the framework lacks explicit estimation of scene dynamics in 3D scene understanding. In future work, we would like to explore methods for modeling scene dynamics through motion segmentation to improve the performance of unsupervised monocular depth estimation in dynamic scenes.

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

Conceptualization, XW; methodology, XW and JS; software, YS; validation, JY, YS, and ZS; formal analysis, XW; writing—original draft preparation, YY; writing—review and editing, XW; supervision, XW and HQ. All authors have read and agreed to the published version of the manuscript.

Funding

This work is supported by the Natural Science Foundation of Chongqing (CSTB2022NSCQ-MSX0398, CSTB2022NSCQ-MSX1425), Science and Technology Foundation of the Education Department of Chongqing (KJQN202101510).

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. Zhao C, Sun Q, Zhang C, Tang Y, Qian F. Monocular depth estimation based on deep learning: An overview. *Sci China Technol Sci* (2020) 63(9):1612–27. doi:10.1007/s11431-020-1582-8

2. Ming Y, Meng X, Fan C, Yu H. Deep learning for monocular depth estimation. *A Review Neurocomputing* (2021) 438:14–33.

3. Liu X, Xue N, Wu T. Learning auxiliary monocular contexts helps monocular 3D object detection. *Proc AAAI Conf Artif Intelligence* (2022) 36:1810–8. doi:10.1609/aaai.v36i2.20074

4. Luo S, Dai H, Shao L, Ding Y. M3dssd: Monocular 3d single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2021). p. 6145–54.

5. Bhattacharyya S, Shen J, Welch S, Chen C. Efficient unsupervised monocular depth estimation using attention guided generative adversarial network. *J Real-Time Image Process* (2021) 18(4):1357–68. doi:10.1007/s11554-021-01092-0

6. Ye X, Fan X, Zhang M, Xu R, Zhong W. Unsupervised monocular depth estimation via recursive stereo distillation. *IEEE Trans Image Process* (2021) 30:4492–504. doi:10.1109/tip.2021.3072215

7. Sun Q, Tang Y, Zhang C, Zhao C, Qian F, Kurths J. Unsupervised estimation of monocular depth and VO in dynamic environments via hybrid masks. *IEEE Trans Neural Networks Learn Syst* (2021) 33(5):2023–33. doi:10.1109/tnnls.2021.3100895

8. Garg R, Bg VK, Carneiro G, Reid I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: *European conference on computer vision*. Cham: Springer (2016). p. 740–56.

9. Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE (2017). p. 270–9.

10. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE (2017). p. 1851–8.

11. Ding X, Guo Y, Ding G, Han J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV); October 2019. IEEE (2019). p. 1911–20.

12. Liu S, Huang D. Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV). IEEE (2018). p. 385–400.

13. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE (2018). p. 7132–41.

14. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); December 2016. IEEE (2016). p. 2921–9.

15. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). IEEE (2017). p. 2881–90.

16. Liu C, Zhu L, Belkin M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Appl Comput Harmonic Anal* (2022) 59:85–116. doi:10.1016/j.acha.2021.12.009

17. Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans robotics* (2015) 31(5):1147–63. doi:10.1109/tro.2015.2463671

18. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. *Adv Neural Inf Process Syst* (2014) 27:2366–74.

19. Liu M, Salzmann M, He X. Discrete-continuous depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2014; Columbus, OH, USA. IEEE (2014). p. 716–23.

20. Li R, Wang S, Long Z, Gu D. Undeepvo: Monocular visual odometry through unsupervised deep learning. In: Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA); May 2018; Brisbane, QLD, Australia. IEEE (2018). p. 7286–91.

21. Yang Z, Wang P, Xu W, Zhao L, Nevatia R. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. *Proc AAAI Conf Artif Intelligence* (2018) 32:12257. doi:10.1609/aaai.v32i1.12257

22. Dovesi PL, Poggi M, Andraghetti L, Martí M, Kjellström H, Pieropan A, Mattoccia S. Real-time semantic stereo matching. In: Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA); Paris, FranceMay 2020. IEEE (2020). p. 10780–7.

23. Kundu JN, Uppala PK, Pahuja A, Babu RV. Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; March 2018. IEEE (2018). p. 2656–65.

24. Chen X, Wang Y, Chen X, Zeng W. S2r-depthnet: Learning a generalizable depth-specific structural representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 2021; Nashville, TN, USA. IEEE (2021). p. 3034–43.

25. Yin Z, Shi J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2018. IEEE (2018). p. 1983–92.

26. Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); June 2018. IEEE (2018). p. 5667–75.

27. Yang Z, Wang P, Wang Y, Xu W, Nevatia R. Lego: Learning edge with geometry all at once by watching videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2018. IEEE (2018). p. 225–34.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 18 January 2023
DOI 10.3389/fphy.2023.1121311

OPEN ACCESS

EDITED BY
Bo Xiao,
Imperial College London, United Kingdom

REVIEWED BY
Guanqiu Qi,
Buffalo State College, United States
Jian Sun,
Southwest University, China

*CORRESPONDENCE
Kaizheng Wang,
kz.wang@foxmail.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 11 December 2022
ACCEPTED 09 January 2023
PUBLISHED 18 January 2023

CITATION
Yang Y, Xu K and Wang K (2023), Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection. *Front. Phys.* 11:1121311. doi: 10.3389/fphy.2023.1121311

Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection

Yang Yang[1], Kaixiong Xu[1] and Kaizheng Wang[2]*

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [2]Faculty of Electrical Engineering, Kunming University of Science and Technology, Kunming, China

Multispectral pedestrian detection is a technology designed to detect and locate pedestrians in Color and Thermal images, which has been widely used in automatic driving, video surveillance, etc. So far most available multispectral pedestrian detection algorithms only achieved limited success in pedestrian detection because of the lacking take into account the confusion of pedestrian information and background noise in Color and Thermal images. Here we propose a multispectral pedestrian detection algorithm, which mainly consists of a cascaded information enhancement module and a cross-modal attention feature fusion module. On the one hand, the cascaded information enhancement module adopts the channel and spatial attention mechanism to perform attention weighting on the features fused by the cascaded feature fusion block. Moreover, it multiplies the single-modal features with the attention weight element by element to enhance the pedestrian features in the single-modal and thus suppress the interference from the background. On the other hand, the cross-modal attention feature fusion module mines the features of both Color and Thermal modalities to complement each other, then the global features are constructed by adding the cross-modal complemented features element by element, which are attentionally weighted to achieve the effective fusion of the two modal features. Finally, the fused features are input into the detection head to detect and locate pedestrians. Extensive experiments have been performed on two improved versions of annotations (sanitized annotations and paired annotations) of the public dataset KAIST. The experimental results show that our method demonstrates a lower pedestrian miss rate and more accurate pedestrian detection boxes compared to the comparison method. Additionally, the ablation experiment also proved the effectiveness of each module designed in this paper.

KEYWORDS

multispectral pedestrian detection, attention mechanism, feature fusion, convolutional neural network, background noise

1 Introduction

Pedestrian detection, parsing visual content to identify and locate pedestrians on an image/video, has been viewed as an essential and central task within the computer vision field and widely employed in various applications, e.g. autonomous driving, video surveillance and person re-identification [1–7]. The performance of such technology has greatly advanced through the facilitation of convolutional neural networks (CNN). Typically, pedestrian detectors take Color images as input and try to retrieve the pedestrian information from

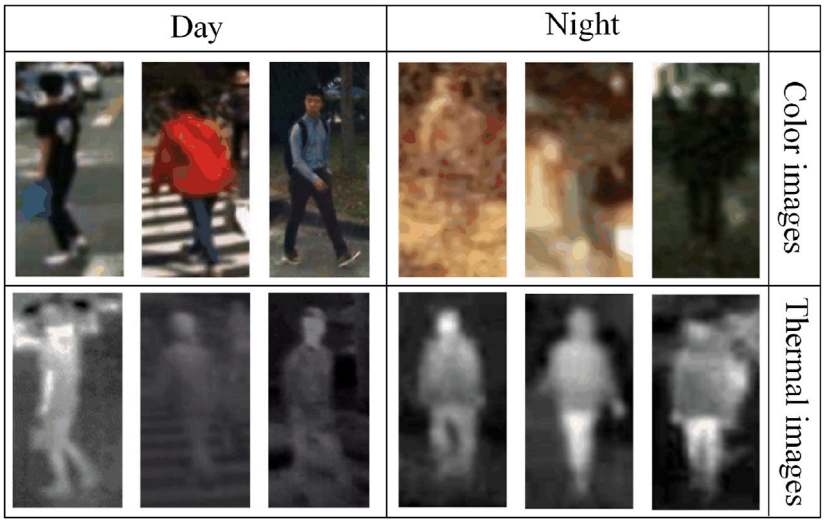


FIGURE 1
Example of color and thermal images of pedestrians in daytime and nighttime scenes.

them. However, the quality of Color images highly depends on the light condition. Missing recognition of pedestrians occurs frequently when pedestrian detectors process Color images with poor resolution and contrast caused by unfavorable lighting. Consequently, the use of such models has been limited for the application of all-weather devices.

Thermal imaging is related to the infrared radiation of pedestrians, barely affected by changes in ambient light. The technique of combining Color and Thermal images has been explored in recent years [8–16]. These methods has been shown to exhibit positive effects on pedestrian detection performance in complex environments as it could retrieve more pedestrian information. However, despite important initial success, there remain two major challenges. First, as shown in Figure 1, the image of pedestrians tends to blend with the background for nighttime Color images resulting from insufficient light [17], and for daytime Thermal images as well due to similar temperatures between the human body and the ambient environment [18]. Second, there is an essential difference between Color images and Thermal images the former displays the color and texture detail information of pedestrians while the latter shows the temperature information. Therefore, solutions needed to be taken to augment the pedestrian features in Color and Thermal modalities in order to suppress background interference, and enable better integration and understanding of both Color and Thermal images to improve the accuracy of pedestrian detection in complex environments.

To address the challenges above, the researches [19,20] designed illumination-aware networks to obtain illumination-measured parameters of Color and Thermal images respectively, which were used as fusion weights for Color and Thermal features in order to realize a self-adaptively fuse of two modal features. However, the acquisition of illumination-measured parameters relied heavily on the classification scores, the accuracy of which was limited by the performance of the classifier. [21] reported confidence-aware networks to predict the confidence of detection boxes for each modal, and then Dempster-Sheffer theory combination rules were employed to fuse the results of different branches based on uncertainty. Nevertheless, the accuracy of predicting the detection boxes' confidence is also affected by the performance of the confidence-aware network. A cyclic fusion and refinement scheme was introduced by [22] for the sake of gradually improving the quality of Color and Thermal features and automatically adjusting the complementary and consistent information balance of the two modalities to effectively utilize the information of both modalities. However, this method only used a simple feature cascade operation to fuse Color and Thermal features and failed to fully exploit the complementary features of these two modalities.

To tackle the problems aforementioned, we propose a multispectral pedestrian detection algorithm with cascaded information enhancement and cross-modal attention feature fusion. The cascaded information enhancement module (CIEM) is designed to enhance the pedestrian information suppressed by the background in the Color and Thermal images. CIEM uses a cascaded feature fusion block to fuse Color and Thermal features to obtain fused features of both modalities. Since the fused features contain the consistency and complementary information of Color and Thermal modalities, the fused features can be used to enhance Color and Thermal features respectively to reduce the interference of background on pedestrian information. Inspired by the attention mechanism, the attention weights of the fused features are sequentially obtained by channel and spatial attention learning, and the Color and Thermal features are multiplied with the attention weights element by element, respectively. In this way, the single-modal features have the combined information of the two modalities, and the single-modal information is enhanced from the perspective of the fused features. Although CIEM enriches single-modal pedestrian features, simple feature fusion of the enhanced single-modal features is still insufficient for robust multispectral pedestrian detection. Thus, we design the cross-modal attention feature fusion module (CAFFM) to efficiently fuse Color and Thermal features. Cross-modal attention is used in this module to implement the differentiation of pedestrian features between different modalities. In order to supplement the pedestrian information of the other modality to the local modality, the attention of the other

modality is adopted to augment the pedestrian features of the local modality. A global feature is constructed by adding the Color and Thermal features after performing cross-modal feature enhancement, and the global feature is used to guide the fusion of the Color and Thermal features. Overall, the method presented in this paper enables more comprehensive pedestrian features acquisition through cascaded information enhancement and cross-modal attention feature fusion, which effectively enhances the accuracy of multispectral image pedestrian detection. The main contributions of this paper are summarized as follows.

(1) A cascaded information enhancement module is proposed. From the perspective of fused features, it reduces the interference from the background of Color and Thermal modalities on pedestrian detection and augments the pedestrian features of Color and Thermal modalities separately through an attention mechanism.
(2) The designed cross-modal attention feature fusion module first mines the features of both Color and Thermal modalities separately through a cross-modal attention network and adds them to the other modality for cross-modal feature enhancement. Meanwhile, the cross-modal enhanced Color and Thermal features are used to construct global features to guide the feature fusion of the two modalities.
(3) Numerous experiments are conducted on the public dataset KAIST to demonstrate the effectiveness and superiority of the proposed method. In addition, the ablation experiments also demonstrate the effectiveness of the proposed modules.

2 Related works

2.1 Multispectral pedestrian detection

Multispectral sensors can obtain paired Color-Thermal images to provide complementary information about pedestrian targets. A large multispectral pedestrian detection (KAIST) dataset was constructed by [8]. Meanwhile, by combining the traditional aggregated channel feature (ACF) pedestrian detector [23] with the HOG algorithm [24], an extended ACF (ACF + T + THOG) method was proposed to fuse Color and Thermal features. In 2016, [9] proposed four fusion modalities of low-layer feature, middle-layer feature, high-layer feature, and confidence fraction fusion with VGG16 as the backbone network, and the middle-layer feature fusion was proved to offer the maximum integration capability of Color and Thermal features. Inspired by this, [25] developed a multispectral region candidate network with Faster RCNN (Region with CNN features, RCNN) [26] as the architecture and replaced the original classifier in Faster RCNN with an enhanced decision tree classifier to reduce the missed and false detection of pedestrians. Recently,[27] deployed the EfficientDet as the backbone network and proposed an EfficientDet-based fusion framework for multispectral pedestrian detection to improve the detection accuracy of pedestrians in Color and Thermal images by adding and cascading the Color and Thermal features. Although the studies [8,9,25,27] fused Color and Thermal features for pedestrian detection, they mainly focused on exploring the impact of different stages of fusion on pedestrian detection, and only adopted simple feature fusion and not focusing on the case of pedestrian and background confusion.

In 2019, [28] observed a weak alignment problem of pedestrian position between Color and Thermal images, for which the KAIST dataset was re-annotated and Aligned Region CNN (AR-CNN) was proposed to handle weakly aligned multispectral pedestrian detection data in an end-to-end manner. But the deployment of this algorithm requires pairs of annotations, and the annotation of the dataset is a time-consuming and labor-intensive task, which makes the algorithm difficult to be applied in realistic scenes. [29] proposed a new single-stage multispectral pedestrian detection framework. This framework used multi-label learning to learn input state-aware features based on the state of the input image pair by assigning an individual label (if the pedestrian is visible in only one image of the image pair, the label vector is assigned as $y_1 \in [0, 1]$ or $y_2 \in [1, 0]$; if the pedestrian is visible in both images of the image pair, the label vector is assigned as $y_3 \in [1, 1]$) to solve the problem of weak alignment of pedestrian locations between Color and Thermal images, but the model still requires pairs of annotations during training. [19] designed illumination-aware networks to obtain illumination-measured parameters for Color and Thermal images separately and used them as the fusion weights for Color and Thermal features. [20] designed a differential modality perception fusion module to guide the features of the two modalities to become similar, and then used the illumination perception network to assign fusion weights to the Color and Thermal features. [30] reported an uncertainty-aware cross-modal guidance (UCG) module to guide the distribution of modal features with high prediction uncertainty to align with the distribution of modal features with low prediction uncertainty. The researches [19,20] noticed that the pedestrians in Color and Thermal images are easily confused with the background and used illumination-aware networks to assign fusion weights to Color and Thermal features. However, the acquisition of illumination-measured parameters relied heavily on the classification scores, whose accuracy was limited by the performance of the classifier. In contrast, the method proposed in this paper not only considers the confusion of pedestrians and background in Color and Thermal images but also effectively fuses the two modal features.

2.2 Attention mechanisms

Attention mechanisms [31] utilized in computer vision are aimed to perform the processing of visual information. Currently, attention mechanisms have been widely used in semantic segmentation [32], image captioning [33], image fusion [34,35], image dehazing [36], saliency target detection [37], person re-identification [38–40], etc. [41] introduced the idea of a squeeze and excitation network (SENet) to simulate the interdependence between feature channels in order to generate channel attention to recalibrate the feature mapping of channel directions. [42] employed the use of a selective kernel unit (SKNet) to adaptively fuse branches with different kernel sizes based on input information. A work inspired by this was from [43]. They designed a multi-scale channel attention feature fusion network that used channel attention mechanisms to replace simple fusion operations such as feature cascades or summations in feature fusion to produce richer feature representations. However, this recent progress in multispectral pedestrian detection has also been limited to two main challenges the interference caused by background and the difference of fundamental characteristics in Color and Thermal images. Therefore, we propose a multispectral pedestrian detection algorithm with cascaded information enhancement and

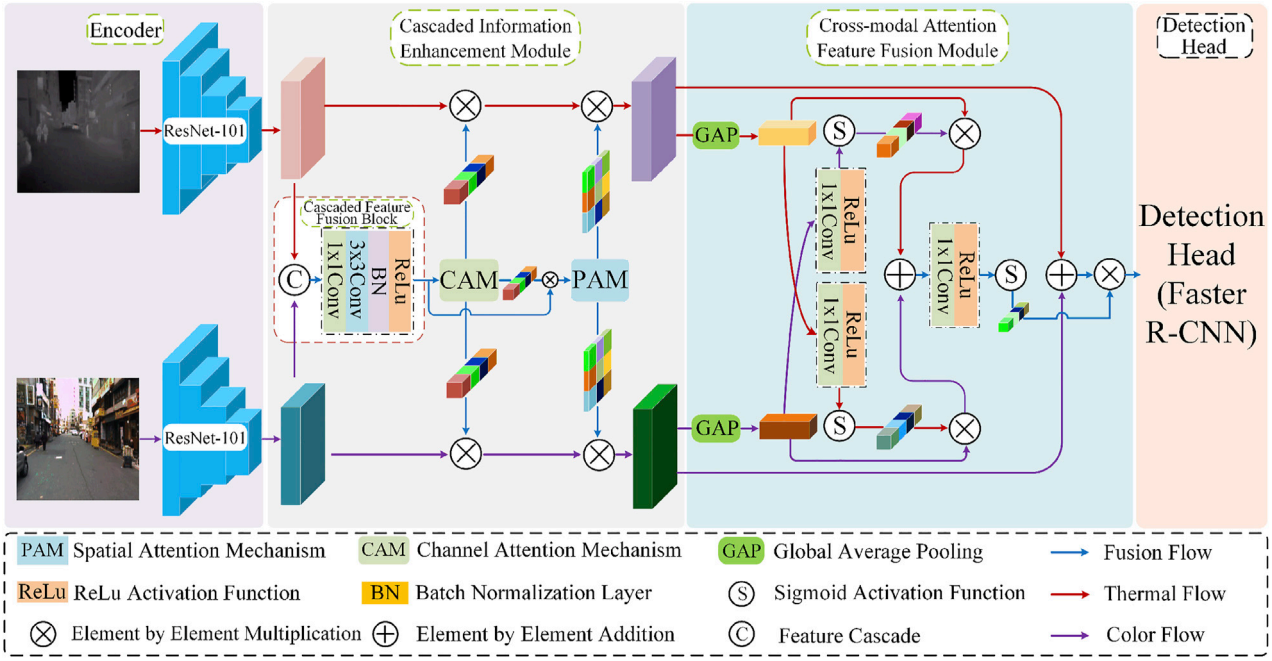


FIGURE 2
Overall framework of the proposed algorithm.

cross-modal attention feature fusion based on the attention mechanism.

3 Methods

The overall network framework of the proposed algorithm is shown in Figure 2. The network consists of an encoder, a cascaded information enhancement module (CIEM), a cross-modal attentional feature fusion module (CAFFM) and a detection head. Specifically, ResNet-101 [44] is used as the backbone network of the encoder to encode the features of the input Color images $\boldsymbol{X}_c$ and Thermal images $\boldsymbol{X}_t$ to obtain the corresponding feature maps $\boldsymbol{F}_c \in \mathrm{R}^{W\times H\times C}$ and $\boldsymbol{F}_t \in \mathrm{R}^{W\times H\times C}$ (W, H, C represent the width, height and the number of channels of the feature maps, respectively). CIEM enhances single-modal information from the perspective of fused features by cascading feature fusion blocks to fuse $\boldsymbol{F}_c$ and $\boldsymbol{F}_t$, and attention weighting the fused features to enrich pedestrian features. CAFFM complements the features of different modalities by mining the complementary features between the two modalities and constructs global features to guide the effective fusion of the two modal features. The detection head is employed for pedestrian recognition and localization of the final fused features.

3.1 Cascaded information enhancement module

Considering the confusion of pedestrians with the backgrounds in Color and Thermal images, we design a cascaded information enhancement module (CIEM) to augment the pedestrian features of both modalities to mitigate the effect of background interference on pedestrian detection. Specifically, a cascaded feature fusion block is used to fuse the Color features $\boldsymbol{F}_c$ and Thermal features $\boldsymbol{F}_t$. The cascaded feature fusion block consists of feature cascade, 1 × 1 convolution, 3 × 3 convolution, *BN* layer, and *ReLu* activation function. The feature cascade operation splice $\boldsymbol{F}_c$ and $\boldsymbol{F}_t$ along the direction of channels. 1 × 1 convolution is conducive to cross-channel feature interaction in the channel dimension and reducing the number of channels in the splice feature map, while 3 × 3 convolution expands the field of perception and makes a more comprehensive fusion of features for generating fusion features $\boldsymbol{F}_{ct}$:

$$\boldsymbol{F}_{ct} = ReLu\left(BN\left(Conv_3\left(Conv_1\left[\boldsymbol{F}_c, \boldsymbol{F}_t\right]\right)\right)\right) \tag{1}$$

where *BN* denotes batch normalization, $Conv_n(\cdot)$ denotes a convolution kernel with kernel size $n \times n$, $[\cdot, \cdot]$ denotes the cascade of features along the channel direction, $ReLu(\cdot)$ represents *ReLu* activation function. Fusion feature $\boldsymbol{F}_{ct}$ is used to enhance the single-modal information because $\boldsymbol{F}_{ct}$ combines the consistency and complementarity of the Color features $\boldsymbol{F}_c$ and Thermal features $\boldsymbol{F}_t$. The use of $\boldsymbol{F}_{ct}$ for enhancing the single-modal feature can reduce the interference of the noise in the single-modal features (for example, it is difficult to distinguish between the pedestrian information and the background noise).

In order to effectively enhance pedestrian features, the fusion feature $\boldsymbol{F}_{ct}$ is sent into the channel attention module (CAM) and spatial attention module (PAM) [45] to make the network pay attention to pedestrian features. The network structure of CAM and PAM is shown in Figure 3. $\boldsymbol{F}_{ct}$ first learns the channel attention weight $\boldsymbol{w}_{ca} \in \mathrm{R}^{1\times 1\times C}$ by CAM, then uses $\boldsymbol{w}_{ca}$ to weight $\boldsymbol{F}_{ct}$, and the spatial attention weight $\boldsymbol{w}_{pa} \in \mathrm{R}^{W\times H\times 1}$ is obtained from the weighted features by PAM.

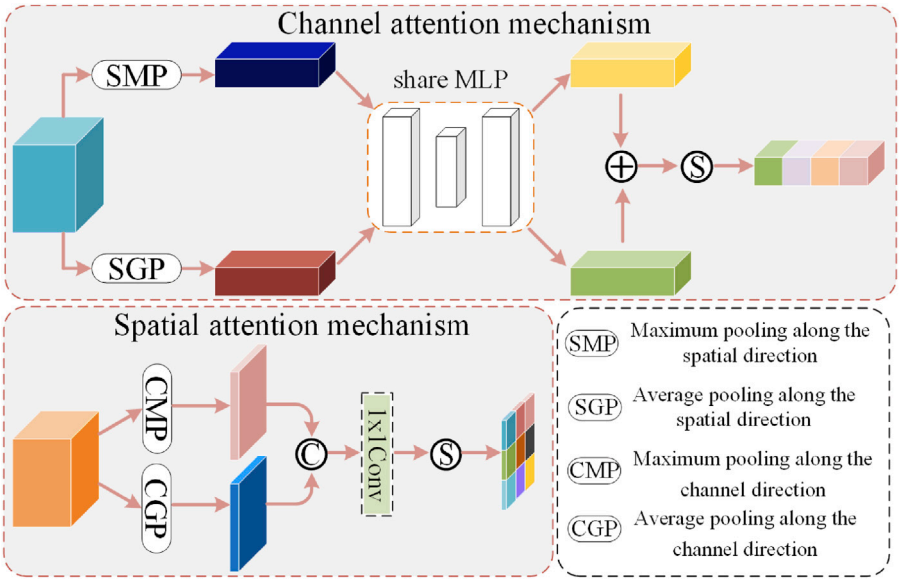


FIGURE 3
Network structure of channel attention and spatial attention.

The single-modal Color features $\boldsymbol{F}_c$ and Thermal features $\boldsymbol{F}_t$ are multiplied element by element with the attention weights $\boldsymbol{w}_{ca}$ and $\boldsymbol{w}_{pa}$ to enhance the single-modal features from the perspective of fused features. The whole process can be described as follows:

$$\boldsymbol{F}_t' = (\boldsymbol{F}_t \otimes \boldsymbol{w}_{ca}) \otimes \boldsymbol{w}_{pa} \tag{2}$$

$$\boldsymbol{F}_c' = (\boldsymbol{F}_c \otimes \boldsymbol{w}_{ca}) \otimes \boldsymbol{w}_{pa} \tag{3}$$

where $\boldsymbol{F}_t'$ and $\boldsymbol{F}_c'$ denote the Color features and Thermal features obtained by the cascaded information enhancement module, respectively. $\otimes$ represents the element by element multiplication.

3.2 Cross-modal attention feature fusion module

There is an essential difference between Color and Thermal images, Color images reflect the color and texture detail information of pedestrians while Thermal images contain the temperature information of pedestrians, however, they also have some complementary information. In order to explore the complementary features of different image modalities and fuse them effectively, we design a cross-modal attention feature fusion module.

Specifically, the Color features $\boldsymbol{F}_c'$ and Thermal features $\boldsymbol{F}_t'$ enhanced by CIEM are first mapped into feature vectors $\boldsymbol{v}_c \in \mathrm{R}^{1\times1\times C}$ and $\boldsymbol{v}_t \in \mathrm{R}^{1\times1\times C}$, respectively, by using global average pooling operation. The cross-modal attention network consists of a set of symmetric 1 × 1 convolutions, *ReLu* activation functions, and *Sigmoid* activation functions. In order to obtain the complementary features of the two modalities, more pedestrian features need to be mined from the single-modal. The feature vectors $\boldsymbol{v}_t$ and $\boldsymbol{v}_c$ are learned to the respective modal attention weights $\boldsymbol{w}_t \in \mathrm{R}^{1\times1\times C}$ and $\boldsymbol{w}_c \in \mathrm{R}^{1\times1\times C}$ by a cross-modal attention network, and then the Color features $\boldsymbol{F}_c'$ are multiplied element by element with the attention weights $\boldsymbol{w}_t$ of the Thermal modality, and the Thermal features $\boldsymbol{F}_t'$ are multiplied element by element with the attention weights $\boldsymbol{w}_c$ of the Color modality to complement the features of the other modality into the present modality. The specific process can be expressed as follows.

$$\boldsymbol{w}_t = Sigmoid\left(ReLu\left(Conv_1\left(GAP\left(\boldsymbol{F}_t'\right)\right)\right)\right) \tag{4}$$

$$\boldsymbol{F}_{ct}' = \boldsymbol{w}_t \otimes GAP\left(\boldsymbol{F}_c'\right) \tag{5}$$

$$\boldsymbol{w}_c = Sigmoid\left(ReLu\left(Conv_1\left(GAP\left(\boldsymbol{F}_c'\right)\right)\right)\right) \tag{6}$$

$$\boldsymbol{F}_{tc}' = \boldsymbol{w}_c \otimes GAP\left(\boldsymbol{F}_t'\right) \tag{7}$$

where $\boldsymbol{F}_{ct}'$ denotes Color features after supplementation with Thermal features, $\boldsymbol{F}_{tc}'$ denotes Thermal features after supplementation with Color features, $GAP(\cdot)$ denotes global average pooling operation, $Conv_1(\cdot)$ denotes convolution with convolution kernel size 1 × 1, $ReLu(\cdot)$ denotes *ReLu* activation operation, and *Sigmoid* $(\cdot)$ denotes *Sigmoid* activation operation.

In order to efficiently fuse the two modal features, the features $\boldsymbol{F}_{ct}'$ and $\boldsymbol{F}_{tc}'$ are subjected to an element by element addition operation to obtain a global feature vector containing Color and Thermal features. Then, the features $\boldsymbol{F}_t'$ and $\boldsymbol{F}_c'$ are added element by element and multiplied with the attention weight $\boldsymbol{w}_{ct}$ of the global feature vector element by element to guide the fusion of Color and Thermal features from the perspective of global features to obtain the final fused feature $\boldsymbol{F}$. The fused feature $\boldsymbol{F}$ is input to the detection head to obtain the pedestrian detection results. The feature fusion process can be expressed as follows:

$$\boldsymbol{w}_{ct} = Sigmoid\left(ReLu\left(Conv_1\left(\boldsymbol{F}_{ct}' \oplus \boldsymbol{F}_{tc}'\right)\right)\right) \tag{8}$$

$$\boldsymbol{F} = \boldsymbol{w}_{ct} \otimes \left(\boldsymbol{F}_t' \oplus \boldsymbol{F}_c'\right) \tag{9}$$

where $\oplus$ denotes element by element addition.

3.3 Loss function

The loss function in this paper is consistent with the literature [26] and uses the Region Proposal Network (RPN) loss function L_{RPN} and Fast RCNN [46] loss function L_{FR} to jointly optimize the network:

$$L = L_{RPN} + L_{FR} \tag{10}$$

Both L_{RPN} and L_{FR} consist of classification loss L_{cls} and bounding box regression loss L_{reg}:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{11}$$

Where, N_{cls} is the number of anchors, N_{reg} is the sum of positive and negative sample number, p_i is the probability that the i-th anchor is predicted to be the target, p_i^* is 1 when the anchor is a positive sample, otherwise it is 0, t_i denotes the bounding box regression parameter predicting the i-th anchor, and t_i^* denotes the GT bounding box parameter of the i-th anchor, $\lambda = 1$.

The difference between the classification loss of RPN network and Fast RCNN network is that the RPN network focuses only on the foreground and background when classifying, so its loss is a binary cross-entropy loss, while the Fast RCNN classification is focused to the target category and is a multi-category cross-entropy loss:

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \tag{12}$$

The bounding box regression loss of RPN network and Fast RCNN network uses Smooth $_{L_1}$ loss:

$$L_{\text{reg}}(t_i, t_i^*) = R(t_i - t_i^*) \tag{13}$$

Where, R denotes Smooth $_{L_1}$ function,

$$\text{Smooth}_{L_1}(x) = \begin{cases} \dfrac{x^2}{2\sigma^2} & \text{if } |x| < \dfrac{1}{\sigma^2} \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{14}$$

The difference between the bounding box regression loss of RPN loss and the regression loss of Fast RCNN loss is that the RPN network is trained when $\sigma = 3$ and the Fast RCNN network is trained when $\sigma = 1$.

4 Experimental results and analysis

4.1 Datasets

This paper evaluates the algorithm performance on the KAIST pedestrian dataset [8], which is composed of 95,328 pairs of Color and Thermal images captured during daytime and nighttime. It is the most widely used multispectral pedestrian detection dataset at present. The dataset is labeled with four categories including person, people, person?, and cyclist. Considering the application areas of multispectral pedestrian detection (e.g., automatic driving), all four categories are treated as positive examples for detection in this paper. To address the problem of the annotation errors and missing annotations in the original annotation of the KAIST dataset, studies [9,28,47] performed data cleaning and re-annotation of the original data. Given that the annotations used in various studies are not consistent, we use 7601 pairs of Color and Thermal images from synthetic annotation (SA) [47] and 8892 pairs of Color and Thermal images from paired annotation (PA) [28] for model training. The test set consists of 2252 pairs of Color and Thermal images, of which 1455 pairs are from the daytime and 797 pairs are from the nighttime. For a fair comparison with other methods, the test experiments were performed according to the reasonable settings proposed in the literature [8].

4.2 Evaluation indexes

In this paper, Log-average Miss Rate (MR) proposed by [48] is employed as an evaluation index and combined with the plotting of the Miss Rate-FPPI curve to assess the effectiveness of the algorithm. The horizontal coordinate of the Miss Rate-FPPI curve indicates the average number of False Positives Per Image (FPPI), and the vertical coordinate represents the Miss Rate (MR), which is expressed as:

$$\text{MissRate} = \frac{FN}{TP + FN} \tag{15}$$

$$FPPI = \frac{FP}{\text{Total}(\text{images})} \tag{16}$$

where FN denotes False Negative, TP denotes True Positive, FP denotes False Positive, the sum of TP and FN is the number of all positive samples, and Total (images) denotes the total number of predicted images. It is worth noting that the lower the Miss Rate-FPPI curve trend, the better the detection performance; the smaller the MR value, the better the detection performance. In order to calculate MR, in logarithmic space, nine points are taken from the horizontal coordinate (limited value range is $[10^{-2}, 10^{0}]$) of Miss Rate-FPPI curve, and then there are nine corresponding vertical coordinates m_1, m_2,...m_9. By averaging these values, MR can be obtained as follows:

$$\text{MR} = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln(m_i)\right] \tag{17}$$

where n is 9.

4.3 Implementation details

In this paper, the deep learning framework pytorch1.7 is adopted. The experimental platform is the ubuntu18.04 operating system and a single NVIDIA GeForce RTX 2080Ti GPU. Stochastic Gradient Descent (SGD) algorithm is used to optimize the network during model training, with momentum value of 0.9, weight attenuation value 5×10^{-4}, and initial learning rate is 1×10^{-3}. The model is iterated for five epochs with the batch size of 4, and the learning rate decay to 1×10^{-4} after the 3rd epoch.

4.4 Experimental results and analysis

4.4.1 Construction of the baseline

This work constructs a baseline algorithm architecture based on ResNet-101 backbone network and Faster RCNN detection head. Simple characteristic fusion (feature cascade, element by element addition and element by element multiplication) of the Color and Thermal features output by the backbone network is carried out in

TABLE 1 Experimental results of baseline under different fusion modes.

Fusion modes	All-weather
feature cascade	14.62
element by element multiplication	14.26
element by element addition	**13.84**

The bold values in highlight the optimal results for this column.

TABLE 3 MRs of different modal inputs.

Input	All-weather	Day	Night
dual-stream Color images	25.37	19.31	31.18
dual-stream Thermal images	17.55	22.81	12.61
Color images + Thermal images	**13.84**	**15.35**	**12.48**

The bold values in highlight the optimal results for this column.

TABLE 2 MRs of different methods on KAIST datasets.

Methods	SA			PA(Color)			PA(Thermal)		
	All-weather	Day	Night	All-weather	Day	Night	All-weather	Day	Night
ACF + T + THOG	41.65	39.18	48.29	41.74	39.30	49.52	41.36	38.74	48.30
Halfway Fusion	25.75	24.88	26.59	25.10	24.29	26.12	25.51	25.20	24.90
CMT_CNN	36.83	34.56	41.82	36.25	34.12	41.21	–	–	–
IAF R-CNN	15.73	14.55	18.26	15.65	14.95	18.11	16.00	15.22	17.56
IATDNN + IAMSS	14.95	14.67	15.72	15.14	14.82	15.87	15.08	15.02	15.20
CIAN	14.12	14.77	11.13	14.64	15.13	12.43	14.68	16.21	9.88
CS-RCNN	11.43	11.86	8.82	–	–	–	–	–	–
IT-MN	14.19	14.30	13.98	–	–	–	–	–	–
DCRD	12.58	13.12	11.65	13.64	13.15	13.98	–	–	–
Ours	10.71	13.09	8.45	11.11	12.85	8.77	10.98	13.07	8.53

three sets of experiments. The fused feature is used as the input of the detection head. In order to ensure the high efficiency of the build baseline algorithm, synthesis annotation is employed to train and test the baseline. The test results are shown in Table 1. The MR values using feature cascade, element by element addition and element by element multiplication in the all-weather scene are 14.62%, 13.84% and 14.26%, respectively. By comparing these three results, it can be seen that the feature element by element addition demonstrates the best performance. Therefore, we adopt the method of adding features element by element as the baseline integration method.

4.4.2 Performance comparison of different methods

The performance of this method is compared with several other state-of-the-art methods. The compared methods include hand-represented methods, e.g., ACT + T + THOG [8] and deep learning-based methods, e.g., Halfway Fusion [9], CMT_CNN[49], CIAN[50], IAF R-CNN[51], IATDNN + IAMSS[19], CS-RCNN [52], IT-MN [53], and DCRD [54]. Here, the model is trained using 7601 pairs of Color and Thermal images from SA and 8892 pairs of Color and Thermal images from PA, respectively. Besides, 2252 pairs of Color and Thermal images from the test set are used for model testing. Table 2 lists the experimental results.

Table 2 shows that when the model is trained with SA, the MRs of the method proposed in this paper are 10.71%, 13.09% and 8.45% for all-weather, daytime and nighttime scenes, respectively, which are 0.72%, −1.23% and 0.37% lower than the compared method CS-RCNN with the best performance, respectively. The PA (Color) and PA (Thermal) in Table 2 represent the Color annotation and Thermal annotation in the pairwise annotation PA, respectively, for the purpose of training the model. It can be seen from two that the MRs of the method in this paper are 11.11% and 10.98% when using Color annotation and Thermal annotation in the all-weather scene, which are 2.53% and 3.70%, respectively, lower than those of compared method with the best performance. In addition, by analyzing the experimental results of two improved versions of annotations, it can be found that pedestrian detection results are different when using different annotations, indicating the importance of annotations.

4.4.3 Analysis of ablation experiments

4.4.3.1 Complementarity and importance of color and thermal features

This section compares the effect of different input sources on pedestrian detection performance. In order to eliminate the impact of the proposed module on detection performance, three sets of experiments are conducted on baseline: 1) the combination of Color and Thermal images as the input source (the input of the two branches of the backbone network are respectively Color and Thermal images); 2) dual-stream Color image as the input source (use Color images to replace Thermal images, that is, the backbone network input source is Color images); 3) dual-stream Thermal images as the input source (use Thermal images to replace Color images, that is, the backbone network input source is Thermal images).The training set of the model here is 7061 pairs of images of SA, and the test set is 2252 pairs of Color and Thermal images. Table 3 shows the MRs of these three input sources for the all-weather, daytime, and nighttime scenes. It can be seen from Table 3 that the MRs obtained using Color and Thermal images as input to the network are

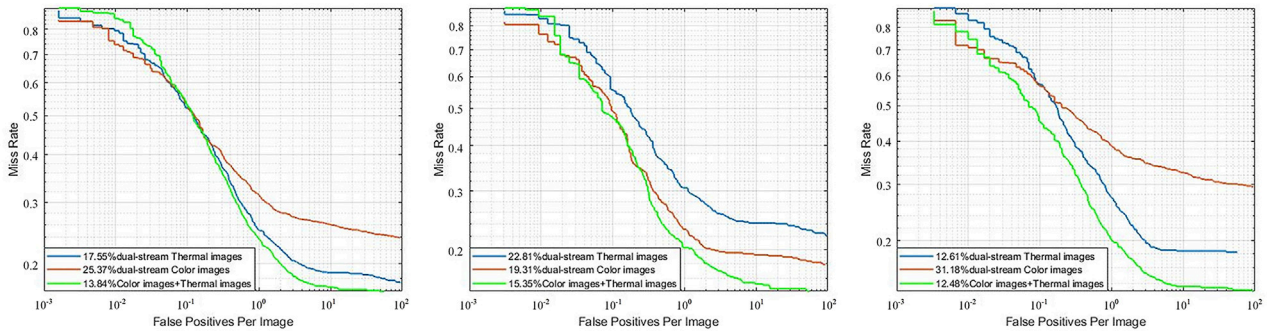


FIGURE 4
The Miss Rate-FPPI curves of the detection results of the three groups of input sources in the All-weather, Daytime and Nighttime scenes (From left to right, All-weather, Daytime and Nighttime Miss Rate-FPPI curves are shown in the figure).

TABLE 4 MRs for ablation studies of the proposed method on SA.

Methods	All-weather	Day	Night
Baseline	13.84	15.35	12.48
baseline + CIEM	11.21	13.15	9.07
baseline + CAFFM	11.68	13.81	9.50
Overall model	**10.71**	**13.09**	**8.45**

The bold values in highlight the optimal results for this column.

13.84%, 15.35% and 12.48% for the all-weather, daytime and nighttime scenes, respectively, which are 11.53%, 3.96%, 18.70% and 3.71%, 7.46%, 0.13% lower than using Color images and Thermal images as input alone. The experimental results prove that the detection network combining Color and Thermal features delivers better performance, indicating that Color and Thermal features are important for pedestrian detection.

Figure 4 shows the Miss Rate-FPPI curves of the detection results for these three input sources in the all-weather, daytime, and nighttime scenes (blue, red and green curves indicate dual-stream Thermal images, dual-stream Color images, and Color and Thermal images, respectively). By analyzing the Miss Rate-FPPI curve trend and combining with the experimental data in Table 3, it can be seen that the detection effect of Color images as the input source is better than that of Thermal images in the daytime scene while the result is the opposite for the night scene, and the detection effect of Color and Thermal images combined as the input source is better than that of single-modal input in both daytime and nighttime. It shows that there are complementary features between Color and Thermal modalities, and the fusion of the two modal features can improve the pedestrian detection performance.

4.4.3.2 Ablation experiments

In this section, ablation experiments are conducted to demonstrate the effectiveness of the proposed cascaded information enhancement module (CIEM) and cross-modal attentional feature fusion module (CAFFM). Here, 7061 pairs of SA images are used to train the model, and 2252 pairs of Color and Thermal images in the test set are used to test the model.

Effectiveness of CIEM: CIEM is used to enhance the pedestrian features in Color and Thermal images to reduce the interference from the background. The experimental results are shown in Table 4. The MRs of baseline on SA are 13.84%, 15.35% and 12.48% for all-weather, daytime and nighttime scenes, respectively. When CIEM is additionally employed, the MRs are 11.21%, 13.15% and 9.07% for all-weather, daytime and nighttime scenes, respectively, which are reduced by 2.63%, 2.20% and 3.41% compared to the baseline, respectively. It is shown that the proposed CIEM effectively enhances the pedestrian features in both modalities, reduces the interference of background, and improves the pedestrian detection performance.

Validity of CAFFM: CAFFM is used to effectively fuse Color and Thermal features. The experimental results are shown in Table 4. On the SA, when the baseline is used with CAFFM, the MRs are 11.68%, 13.81% and 9.50% in all-weather, daytime and nighttime scenes, respectively, which are reduced by 2.16%, 1.54% and 2.98% compared baseline, respectively. It shows that the proposed CAFFM effectively fuses the two modal features to achieve robust multispectral pedestrian detection.

Overall effectiveness: The proposed CIEM and CAFFM are additionally used on the basis of baseline. Experimental results show a reduction of 3.13%, 2.26% and 4.03% in MRs for all-weather, daytime and nighttime scenes, respectively, compared to the baseline, indicating the overall effectiveness of the proposed method. A closer look reveals that with additional employment of CIEM and CAFFM alone, MRs are decreased by 2.63% and 2.16%, respectively, in the all-weather scene, but the MR of the overall model is reduced by 3.13%. It demonstrates that there is some orthogonal complementarity in the role of the proposed two modules.

Figure 5 shows the Miss Rate-FPPI curves for CIEM and CAFFM ablation studies in all-weather, daytime and nighttime scenes (blue, red, orange and green curves represent baseline, baseline + CIEM, baseline + CAFFM and overall model, respectively). It is clear that the curve trends of each module and the overall model are both lower than that of the baseline, which further proves the effectiveness of the method presented in this work.

Furthermore, in order to qualitatively analyze the effectiveness of the proposed CIEM and CAFFM, four pairs of Color and Thermal images (two pairs of images are taken from daytime and two pairs of images are taken from nighttime) are selected from the test set for testing. The pedestrian detection results of the baseline and each proposed module are shown in Figure 6. The first row is the

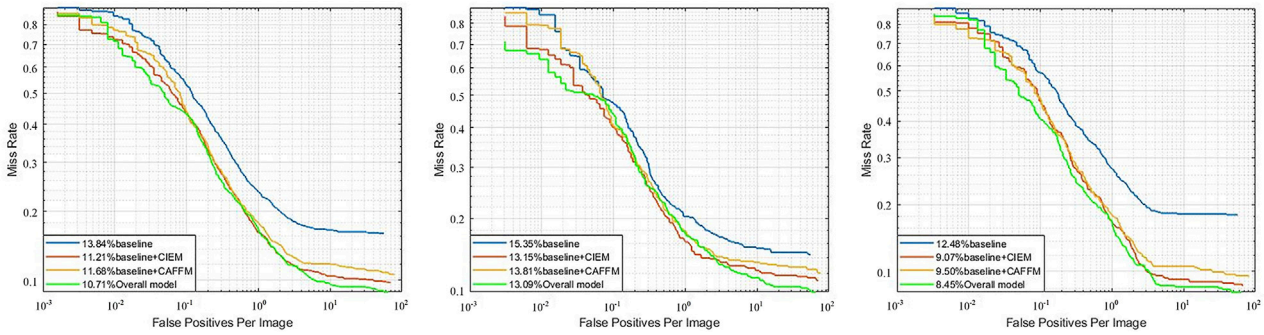


FIGURE 5
The Miss Rate-FPPI curves of CIEM and CAFFM ablation studies in All-weather, Daytime and Nighttime scenes (From left to right, All-weather, Daytime and Nighttime Miss Rate-FPPI curves are shown in the figure).

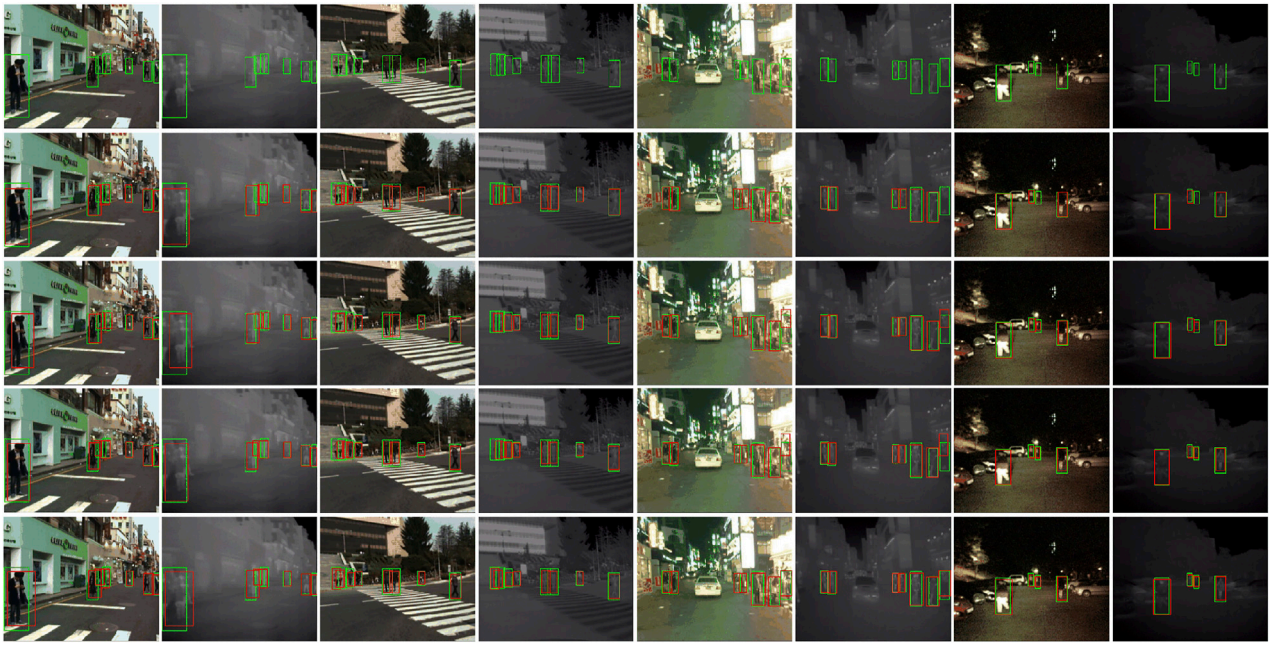


FIGURE 6
In this paper, each module and baseline pedestrian detection results (The first row is the visualization results of labeled boxes for Color and Thermal images, and the second to the fifth rows are the visualization results of the labeled and prediction boxes for baseline, baseline + CIEM, baseline + CAFFM and the overall model pedestrian detection with the green and red boxes representing the labeled and prediction boxes, respectively).

visualization results of labeled boxes for Color and Thermal images, and the second to the fifth rows are the visualization results of the labeled and prediction boxes for baseline, baseline + CIEM, baseline + CAFFM, and the overall model pedestrian detection with the green and red boxes representing the labeled and prediction boxes, respectively. It can be seen that the proposed method successfully addresses the problem of pedestrian missing detection in complex environments and achieves more accurate detection boxes. For example, the second row, pedestrian detection missing happens in the first, third, and fourth pairs of images in the baseline detection result, however, the pedestrian miss detection problem is properly solved with CIEM and CAFFM added to the baseline and the overall model produces more accurate pedestrian detection boxes.

5 Conclusion

In this paper, we propose a multispectral pedestrian detection algorithm including cascaded information enhancement module and cross-modal attention feature fusion module. The proposed method improves the accuracy of pedestrian detection in multispectral images (Color and Thermal images) by effectively fusing the features from the two modules and augmenting the pedestrian features. Specifically, on the one hand, a cascaded information enhancement module (CIEM) is designed to enhance single-modal features to enrich the pedestrian features and suppress interference from the background noise. On the other hand, unlike previous methods that simply splice Color and Thermal features directly, a cross-modal attention feature fusion

module (CAFFM) is introduced to mine the features of both Color and Thermal modalities and to complement each other, then complementary enhanced modal features are used to construct global features. Extensive experiments have been conducted on two improved annotations of the public dataset KAIST. The experimental results show that the proposed method is conducive to obtain more comprehensive pedestrian features and improve the accuracy of multispectral image pedestrian detection.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://gitcode.net/mirrors/soonminhwang/rgbt-ped-detection?utm_source=csdn_github_accelerator.

Author contributions

YY responsible for scheme design, experiment and writing of the paper. KX guide the scheme design and experiment of the paper. KW guide experimental data analysis, paper writing and modification.

Funding

This work was supported by the National Natural Science Foundation of China (No. 52107017) and Fundamental Research Fund of Science and Technology Department of Yunnan Province(No.202201AU070172).

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. Jeong M, Ko BC, Nam J-Y. Early detection of sudden pedestrian crossing for safe driving during summer nights. *IEEE Trans Circuits Syst Video Technol* (2017) 27:1368–80. doi:10.1109/TCSVT.2016.2539684

2. Zhang S, Cheng D, Gong Y, Shi D, Qiu X, Xia Y, et al. Pedestrian search in surveillance videos by learning discriminative deep features. *Neurocomputing* (2018) 283:120–8. doi:10.1016/j.neucom.2017.12.042

3. Li L, Xie M, Li F, Zhang Y, Li H, Tan T. Unsupervised domain adaptive person re-identification guided by low-rank priori. *J Chongqing Univ* (2021) 44:57–70. doi:10.11835/j.issn.1000-582X.2021.11.008

4. Li H, Chen Y, Tao D, Yu Z, Qi G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Trans Inf Forensics Security* (2021) 16:1480–94. doi:10.1109/TIFS.2020.3036800

5. Li H, Dong N, Yu Z, Tao D, Qi G. Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. *IEEE Trans Circuits Syst Video Technol* (2022) 32:2814–30. doi:10.1109/TCSVT.2021.3099943

6. Li S, Li F, Wang K, Qi G, Li H. Mutual prediction learning and mixed viewpoints for unsupervised-domain adaptation person re-identification on blockchain. *Simulation Model Pract Theor* (2022) 119:102568. doi:10.1016/j.simpat.2022.102568

7. Wang S, Liu R, Li H, Qi G, Yu Z. Occluded person re-identification via defending against attacks from obstacles. *IEEE Trans Inf Forensics Security* (2023) 18:147–61. doi:10.1109/TIFS.2022.3218449

8. Hwang S, Park J, Kim N, Choi Y, Kweon IS. Multispectral pedestrian detection: Benchmark dataset and baseline. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 07-12 June 2015; Boston, MA, USA (2015). p. 1037–45. doi:10.1109/CVPR.2015.7298706

9. Liu J, Zhang S, Wang S, Metaxas DN. Multispectral deep neural networks for pedestrian detection. In: Proceedings of the British Machine Vision Conference 2016; 19-22 September 2016; York, UK (2016).

10. 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

11. Zhang Y, Yang M, Li N, Yu Z. Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion. *Signal Process.* (2020) 167:107327. doi:10.1016/j.sigpro.2019.107327

12. 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

13. Li H, He X, Tao D, 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

14. Li H, Wang Y, 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) 69:1082–102. doi:10.1109/tim.2019.2912239

15. Xie M, Wang J, Zhang Y. A unified framework for damaged image fusion and completion based on low-rank and sparse decomposition. *Signal Processing: Image Commun* (2021) 29:116400. doi:10.1016/j.image.2021.116400

16. Wang S, Huang B, Li H, Qi G, Tao D, Yu Z. Key point-aware occlusion suppression and semantic alignment for occluded person re-identification. *Inf Sci* (2022) 606:669–87. doi:10.1016/j.ins.2022.05.077

17. Zhu Z, Luo Y, Chen S, Qi G, Mazur N, Zhong C, et al. Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification. *J Vis Commun Image Representation* (2021) 80:103303. doi:10.1016/j.jvcir.2021.103303

18. Yang X, Qian Y, Zhu H, Wang C, Yang M. Baanet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. In: 2022 International Conference on Robotics and Automation (ICRA); 23-27 May 2022; Philadelphia, PA, USA (2022). p. 2920–6. doi:10.1109/ICRA46639.2022.9811999

19. 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

20. Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In: *European conference on computer vision*. Berlin, Germany: Springer (2020). p. 787–803.

21. Li Q, Zhang C, Hu Q, Fu H, Zhu P. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Trans Multimedia* (2022) 1. doi:10.1109/tmm.2022.3160589

22. Zhang H, Fromont E, Lefevre S, Avignon B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International Conference on Image Processing (ICIP); 25-28 October 2020; Abu Dhabi, United Arab Emirates (2020). p. 276–80. doi:10.1109/ICIP40778.2020.9191080

23. 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

24. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 20-25 June 2005; San Diego, CA, USA (2005). p. 886–93. doi:10.1109/CVPR.2005.177

25. König D, Adam M, Jarvers C, Layher G, Neumann H, Teutsch M. Fully convolutional region proposal networks for multispectral person detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 21-26 July 2017; Honolulu, HI, USA (2017). p. 243–50. doi:10.1109/CVPRW.2017.36

26. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Machine Intelligence* (2017) 39:1137–49. doi:10.1109/TPAMI.2016.2577031

27. Kim J, Park I, Kim S. A fusion framework for multi-spectral pedestrian detection using efficientdet. In: 2021 21st International Conference on Control, Automation and Systems (ICCAS); 12-15 October 2021; Jeju, Korea (2021). p. 1111–3. doi:10.23919/ICCAS52745.2021.9650057

28. Zhang L, Zhu X, Chen X, Yang X, Lei Z, Liu Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 27 October 2019 - 02 November 2019; Seoul, Korea (South) (2019). p. 5126–36. doi:10.1109/ICCV.2019.00523

29. Kim J, Kim H, Kim T, Kim N, Choi Y. Mlpd: Multi-label pedestrian detector in multispectral domain. *IEEE Robotics Automation Lett* (2021) 6:7846–53. doi:10.1109/LRA.2021.3099870

30. Kim JU, Park S, Ro YM. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Trans Circuits Syst Video Technol* (2022) 32:1510–23. doi:10.1109/TCSVT.2021.3076466

31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach California USA. Curran Associates Inc. (2017).

32. Li S, Zou C, Li Y, Zhao X, Gao Y. Attention-based multi-modal fusion network for semantic scene completion. *Proc AAAI Conf Artif Intelligence* (2020) 34:11402–9. doi:10.1609/aaai.v34i07.6803

33. Li B, Zhou Y, Ren H. Image emotion caption based on visual attention mechanisms. In: 2020 IEEE 6th International Conference on Computer and Communications (ICCC); 11-14 December 2020; Chengdu, China. IEEE (2020). p. 1456–60.

34. 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

35. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

36. Li H, Gao J, Zhang Y, Xie M, Yu Z. Haze transfer and feature aggregation network for real-world single image dehazing. *Knowledge-Based Syst* (2022) 251:109309. doi:10.1016/j.knosys.2022.109309

37. Xu M, Fu P, Liu B, Li J. Multi-stream attention-aware graph convolution network for video salient object detection. *IEEE Trans Image Process* (2021) 30:4183–97. doi:10.1109/TIP.2021.3070200

38. Li H, Xu K, Li J, Yu Z. Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification. *Knowledge-Based Syst* (2022) 251:109315. doi:10.1016/j.knosys.2022.109315

39. Zhang Y, Wang Y, Li H, Li S. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In: MM '22. Proceedings of the 30th ACM International Conference on Multimedia; 10 October 2022; New York, NY, USA. Association for Computing Machinery (2022). p. 3347–55. doi:10.1145/3503161.3548224

40. Wang Y, Qi G, Li S, Chai Y, Li H. Body part-level domain alignment for domain-adaptive person re-identification with transformer framework. *IEEE Trans Inf Forensics Security* (2022) 17:3321–34. doi:10.1109/TIFS.2022.3207893

41. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Machine Intelligence* (2020) 42:2011–23. doi:10.1109/TPAMI.2019.2913372

42. Li X, Wang W, Hu X, Yang J. Selective kernel networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15-20 June 2019; Long Beach, CA, USA (2019). p. 510–9. doi:10.1109/CVPR.2019.00060

43. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 03-08 January 2021; Waikoloa, HI, USA (2021). p. 3559–68. doi:10.1109/WACV48630.2021.00360

44. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27-30 June 2016; NV, USA (2016). p. 770–8. doi:10.1109/CVPR.2016.90

45. Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV) (2018). p. 3–19.

46. Girshick R. Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV); 07-13 December 2015; Santiago, Chile (2015). p. 1440–8. doi:10.1109/ICCV.2015.169

47. Li C, Song D, Tong R, Tang M (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*

48. Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans Pattern Anal Machine Intelligence* (2012) 34:743–61. doi:10.1109/TPAMI.2011.155

49. Xu D, Ouyang W, Ricci E, Wang X, Sebe N. Learning cross-modal deep representations for robust pedestrian detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 26 Jul 2017; Hawaii (2017). p. 4236–44. doi:10.1109/CVPR.2017.451

50. 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

51. 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

52. 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

53. 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* (2022) 9:1282–95. doi:10.1109/TNSE.2021.3139335

54. Liu T, Lam K-M, Zhao R, Qiu G. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Trans Circuits Syst Video Technol* (2022) 32:315–29. doi:10.1109/TCSVT.2021.3060162

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 03 February 2023
DOI 10.3389/fphy.2023.1136021

OPEN ACCESS

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

REVIEWED BY
Puhong Duan,
Hunan University, China
Guanqiu Qi,
Buffalo State College, United States

*CORRESPONDENCE
Kaizheng Wang,
kz.wang@foxmail.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 02 January 2023
ACCEPTED 23 January 2023
PUBLISHED 03 February 2023

CITATION
Zhou F, Wen G, Ma Y, Wang Y, Ma Y, Wang G, Pan H and Wang K (2023), Multilevel feature cooperative alignment and fusion for unsupervised domain adaptation smoke detection.
Front. Phys. 11:1136021.
doi: 10.3389/fphy.2023.1136021

Multilevel feature cooperative alignment and fusion for unsupervised domain adaptation smoke detection

Fangrong Zhou[1], Gang Wen[1], Yi Ma[1], Yifan Wang[1], Yutang Ma[1], Guofang Wang[1], Hao Pan[1] and Kaizheng Wang[2]*

[1]Joint Laboratory of Power Remote Sensing Technology, Electric Power Research Institute, Yunnan Power Grid Company Ltd., China Southern Power Grid, Kunming, China, [2]Faculty of Electrical Engineering, Kunming University of Science and Technology, Kunming, China

Early smoke detection using Digital Image Processing technology is an important research field, which has great applications in reducing fire hazards and protecting the ecological environment. Due to the complex changes of color, shape and size of smoke with time, it is challenging to accurately recognize smoke from a given image. In addition, limited by domain shift, the trained detector is difficult to adapt to the smoke in real scenes, resulting in a sharp drop in detection performance. In order to solve this problem, an unsupervised domain adaptive smoke detection algorithm rely on Multilevel feature Cooperative Alignment and Fusion (MCAF) was proposed in this paper. Firstly, the cooperative domain alignment is performed on the features of different scales obtained by the feature extraction network to reduce the domain difference and enhance the generalization ability of the model. Secondly, multilevel feature fusion modules were embedded at different depths of the network to enhance the representation ability of small targets. The proposed method is evaluated on multiple datasets, and the results show the effectiveness of the method.

KEYWORDS

smoke detection, unsupervised domain adaptive object detection, domain alignment, small object detection, feature fusion

1 Introduction

Natural disasters have always been the main cause of power grid failures. Among them, forest fires are easy to cause serious failures of multiple transmission lines due to coupling, causing irreparable losses to power equipment, posing a great threat to the safe operation of the power system, and even affecting people's normal life. The early occurrence of wildfire is often accompanied by the rise of smoke. Therefore, smoke detection is an important method to effectively avoid fire hazards.

Thanks to the rapid development of deep learning [1–6] and the wide applications in other computer vision tasks such as image fusion [7], image dehazing [8] and semantic segmentation [9], the performances of smoke detection have been remarkably improved in recent year, there are still many difficulties to detect smoke in real time. Usually, the training of deep learning models requires a large amount of data, it is extremely difficult to collect thousands of smoke images and manually label in actual scenes. Some researchers have proposed synthetic smoke datasets [10] to make up for this defect. However, due to the domain gap between the synthetic smoke and the real scene smoke, the performance of the detection model is limited. Figure 1

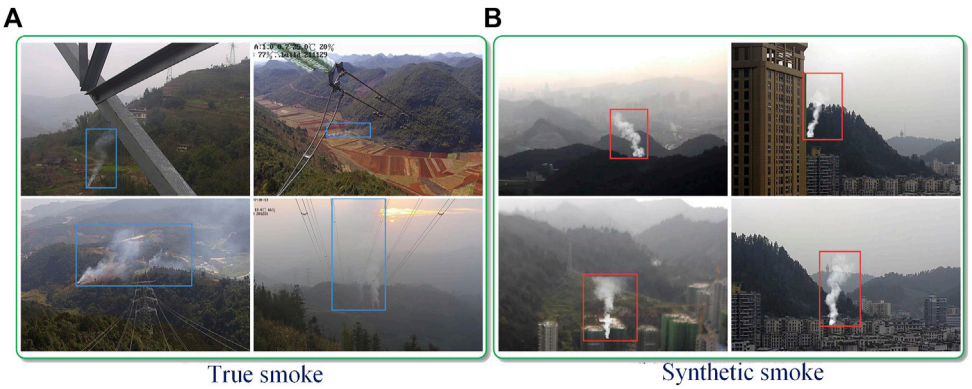


FIGURE 1
(A) Real smoke and **(B)** synthetic smoke. Real smoke has variable color and shape, while synthetic smoke [10] has relatively fixed shape and color.

shows the synthetic smoke and real smoke, there is a large difference between them. In the real scene, due to the unknown weather conditions, it is easy to cause the color, shape and transparency of smoke to change, so that the smoke obtained by the acquisition equipment has differences in resolution, view and brightness, which further increases the difficulty of real-time detection of smoke and fire. It is necessary to study an algorithm that can transfer the knowledge learned from a labeled dataset (source domain) to another unlabeled dataset (target domain).

One approach to solve this problem is Unsupervised Domain Adaptive Object Detection (UDAOD) [11], which aims to adapt the detector using labeled source data and unlabeled target data to alleviate the performance degradation by learning a feature representation that is not affected by domain gap. Existing UDAOD methods can be classified into: style transfer based methods [12–14], self-training based methods [15,16], and domain alignment based methods [17–19].

The method based on style transfer usually uses GAN [20] to transfer the style of the target domain image to the source domain image, and then uses the transformed image to supervise training the detection network, to reduce the domain shift caused by the style difference. However, the smoke image obtained in the real scene has complex background, the image generated by style transfer is different from the real image to some extent, and GAN increases the calculation amount of the model, the final detection performance is highly dependent on the quality of the generated image. Therefore, such methods cannot be well applied to the smoke detection task in the actual scene.

The method based on self-training generally trains the detection model with the source data, then inputs the target data to predict pseudo-label, and finally fine-tunes the model with the pseudo-labels. However, the shape, color and background of smoke in the real scene are not fixed, which is easy to make the predicted pseudo-labels have noise. Fine-tuning the model with noisy pseudo-labels will reduce the detection performance of the model.

Domain alignment based methods achieve feature alignment by adversarial learning. Although such methods have achieved considerable improvement, they align the boundary distribution of the two domains without considering the category information, which may lead to incorrect alignment of samples from different categories of the source domain and the target domain, thus failing to train the best model. The detection category of this work is only smoke, so the above problem does not exist. Existing methods consider the alignment of global features, while this paper aligns features of different scales.

The existing smoke detection work [21,22] pays little attention to cross-domain detection. In order to solve the problems faced by smoke detection in real scenes, this paper proposes a domain adaptive smoke detection algorithm based on multi-level feature fusion and alignment. Specifically, considering the problem of small and fuzzy smoke caused by long shooting distance, the algorithm proposes a multi-level and multi-scale feature fusion strategy to enhance the feature representation ability of the model for small targets. In addition, in order to reduce the domain difference, the algorithm proposes a multi-level feature alignment strategy to reduce the distribution difference between the source domain and the target domain on different levels of features. Compared with the two-stage object detection method, the proposed method is based on YOLOv5 [23], which does not require candidate box prediction and screening, and improves the detection speed.

The main contributions of this paper contain:

- A Multilevel Feature Cooperative domain Alignment method is proposed to reduce the data distribution difference between the source domain and the target domain at the multi-scale feature level.
- A Multilevel Feature Fusion method is proposed to enhance the feature representation ability of small target smoke by fusing features of different scales at different levels of the network.
- The proposed method can perform end-to-end training and detection without additional candidate box calculation and screening, which ensures the training and detection efficiency of the model.

2 Related work

2.1 Object detection

Object detection is a task to classify and locate objects for a given image, which is one of the important research contents in Computer Vision. Recently, deep learning based methods can be divided into two-stage object detection and single-stage object detection.

The two-stage object detection algorithm include two steps: the first step generates the candidate regions, and the second step classifies the candidate regions and regress their positions. The basic idea is to generate regions with high recall such that all objects on the image belong to at least one candidate region. In the second step, the candidate regions generated in the first step are classified by a deep model. Typical two-stage object detection algorithms include R-CNN [24], SPP-net [25], Fast R-CNN [26], Faster R-CNN [27], etc. Due to the large number of candidate regions generated by these algorithms, there are more repeated information and more invalid regions, which leads to large amount of calculation and slow detection speed.

Because the two-stage method needs to process a large number of candidate regions in turn, the detection speed is generally slow. To solve this problem, the one-stage object detection algorithm came into being. Compared with the two-stage object detection algorithm, the one-stage object detection algorithm does not need to generate candidate regions, and directly returns the object category and location on the input image, so the detection speed is faster, but the accuracy is slightly worse. Typical one-stage object detection algorithms include YOLO [28], SSD [29], etc.

2.2 Domain adaptation for object detection

Domain adaptation has been widely studied in Computer Vision. The object detection method based on deep learning is affected by the domain shift, and the network trained on one dataset often performs poorly on other datasets, which is often encountered in real scenarios. Unsupervised Domain Adaptive Object Detection (UDAOD) [30] aims to reduce the domain gap between training data and test data and improve detection performance. Existing UDAOD methods can be divided into: style transfer based methods, self-training based methods, and domain alignment based methods.

Object detection based on style transfer is a popular method in the past few years, and the representative literatures of this kind of method are [12–14]. Hsu et al. [12] proposed a progressive domain adaptation method, which decomposed the problem into two subtasks. Firstly, based on Cycle-GAN [31], synthesized an intermediate domain located in the distribution of the source domain and the target domain, and then adopted a progressive adaptation strategy to gradually narrow the domain gap through the intermediate domain. Inoue et al. [13] believe that the

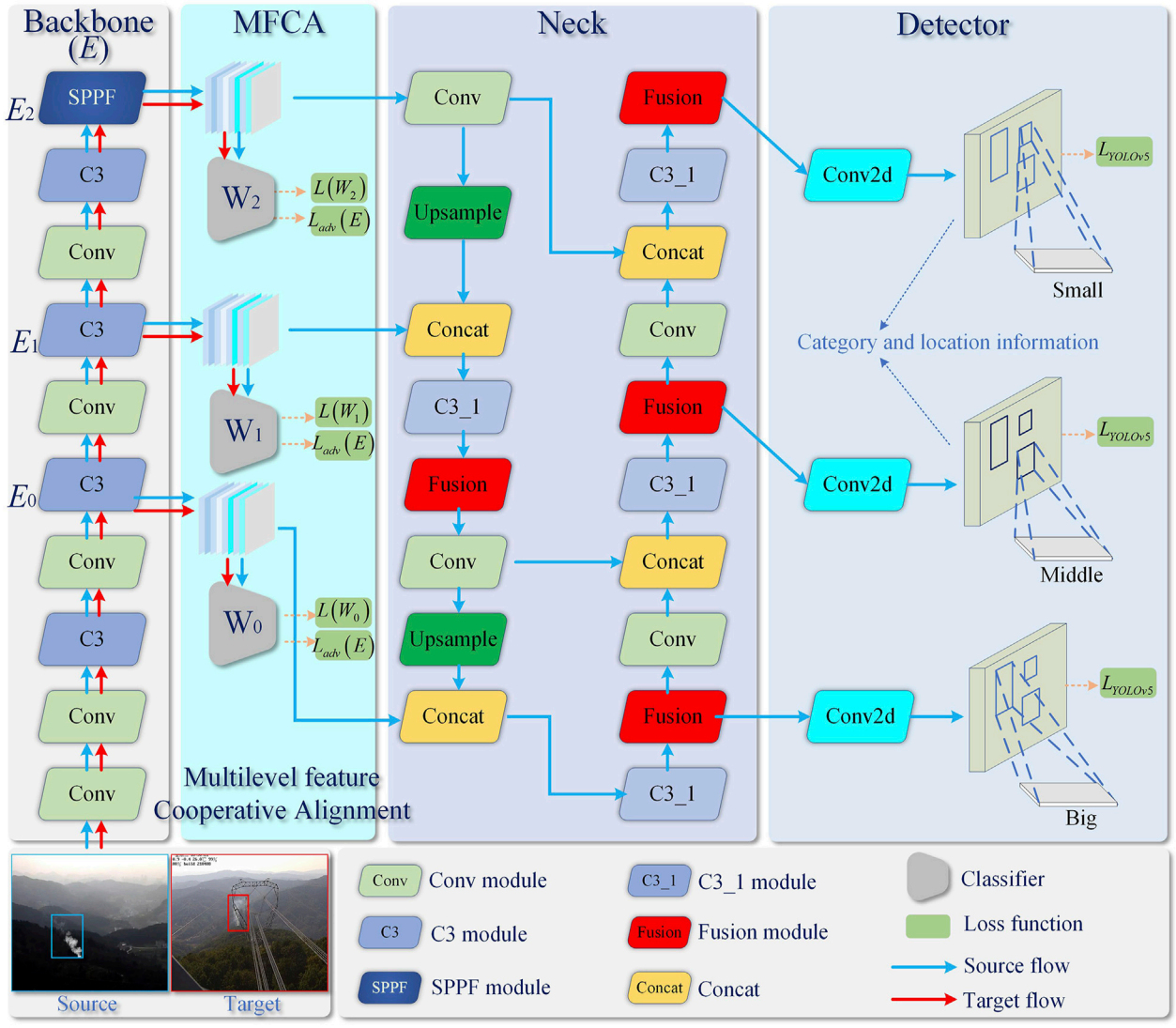


FIGURE 2
The pipeline of our proposed method. See the experiment section for more network details.

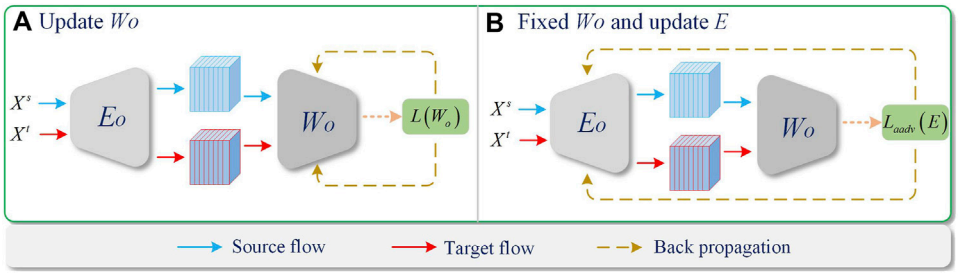


FIGURE 3
Training process. **(A)** Update W_o, **(B)** Fixed W_o and update E.

differences between the source and target domains mainly lie in their underlying features, such as color and texture. By generating similar images with the target domain images based on Cycle-GAN to capture these differences, and then fine-tuning the fully supervised trained detector use the generated images to make the detector robustly to these differences. Kim et al. [14] proposed a two-stage method of Domain Diversification and Multi-domain Invariant Representation Learning to alleviate pixel-level and feature-level domain differences at the same time. In the Domain Diversification stage, samples with different domain differences are generated from labeled source domain data to improve the adaptability of the model. Such methods alleviate the impact of domain differences to a certain extent, but the introduced GAN network increases the amount of computation, and the accuracy of the detector highly depends on the quality of the generated image, which is not suitable for real scenes.

Self-training based object detection [15,16] generally predicts pseudo-labels in the predicted target domain, and then uses the predicted pseudo-labels to fine-tune the model. However, the noise contained in the pseudo-labels will have a negative impact on the performance of the model. Kim et al. [15] proposed a weak self-training method to reduce the adverse effects of inaccurate pseudo-labels and stabilize the learning process. RoyChowdhury et al. [16] proposed an improved knowledge distillation loss by using existing high-confidence detectors to directly obtain the pseudo-labels of the target domain, and studied several methods to assign soft labels to the training samples of the target domain.

Object detection based on domain alignment [17–19] is one of the more commonly used methods. Wu et al. [17] proposed a disentangled representation method based on vector decomposition, attempting to disentangle the representation of domain-invariant features and domain-specific features, to realize domain alignment. Saito et al. [18] proposed a weakly alignment model, which uses adversarial learning to focus the adversarial alignment loss on globally similar images and pays less attention to globally dissimilar images. Zhu et al. [19] believe that the traditional domain adaptive method to align the whole image, while the object detection essentially focuses on the region of interest (local region), and propose to only focus on the relevant region and perform domain alignment.

2.3 Smoke detection

In recent years, researchers have proposed many smoke detection algorithms based on deep learning. Yin et al. [32] proposed a convolutional neural network with depth normalization to automatically extract smoke features and classify them, which reduced the influence of smoke shape and color to a certain extent. In order to further solve the problem of smoke shape and color changes, Gu et al. [33] proposed a dual-channel neural network by successively connecting multiple convolutional layers and Max pooling layers. A batch normalization layer is then selectively attached to each convolutional layer to prevent overfitting and speed up training. zhao et al. [34] proposed that the depth-wise separable method with fixed convolution kernel instead of training iteration was used for smoke detection, which could improve the speed of the algorithm and meet the requirements of real-time fire propagation for speed detection [21]. proposed a Convolutional Neural Network (CNN) -based smoke detection and segmentation framework for clear and hazy environments, employing EfficientNet for better smoke detection [35]. proposed a smoke detection method in normal and foggy weather that combines attention mechanism with feature-level and decision-level fusion modules. An attention mechanism module combining spatial attention and channel attention was proposed to solve the problem of small smoke detection. Secondly, a lightweight feature-level and decision-level fusion module is proposed, which can not only improve the recognition ability of similar objects such as smoke and fog, but also ensure the real-time performance of the model. Zhan et al. [36] proposed a recursive pyramid network with deconvolution and dilated convolution to solve the problem of low detection accuracy caused by high smoke transparency and unclear edges.

Most of the existing smoke detection methods are trained and tested on the same dataset, combined with practical application, this paper focuses on smoke detection in real scenes, and proposes a domain adaptive smoke detection algorithm based on Multilevel feature Cooperative Alignment and Fusion.

3 Proposed method

The structure of Multilevel feature Cooperative Alignment and Fusion (MCAF) algorithm is shown in Figure 2. The algorithm takes YOLOv5 [23] object detection network as the baseline, and is composed of Multilevel Feature Cooperative Alignment module (MFCA) and Multilevel Feature Fusion module (MFF). Specifically, for the given smoke image in the source domain and the target domain, Backbone (denoted as E) is used to extract smoke-related features, and then domain alignment is achieved through the cooperation between multi-scale classifiers W_0, W_1, W_2 in MFCA

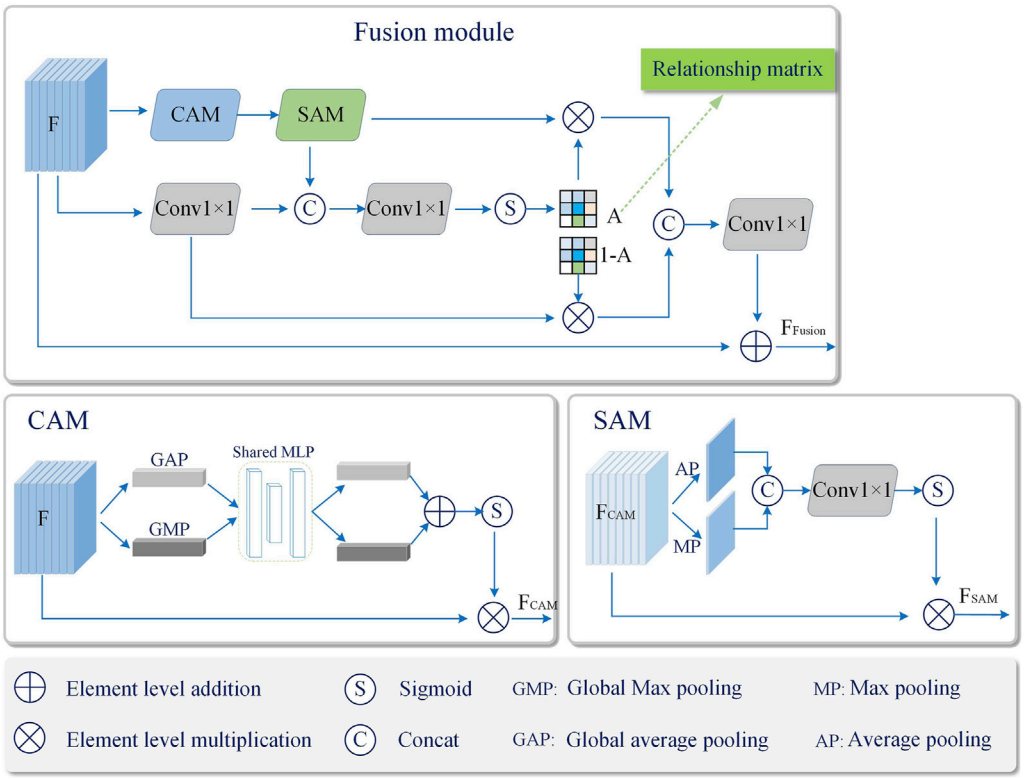


FIGURE 4
The structure of Fusion model.

to reduce the domain differences between the source and target domain features, where, W_0, W_1, W_2 has the same structure, which are consists of a global average pooling, fully connected layer. The difference is that their input feature sizes are not the same. Finally, in the Neck of the detector network (The object detectors developed in recent years often insert some layers between the backbone and the detector, people usually call this part the Neck of the detector) embed feature fusion module at different positions to make the obtained features better adapt to targets of different sizes. Through the end-to-end training of MCAF algorithm, a better detection effect is obtained.

3.1 Model pretrain

In order to make the smoke detection network adapt to real scenarios, pre-training is carried out first. The smoke data in the source and target domains are defined as $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{N^s}$ and $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{N^t}$, where $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ denote the ith sample of the training set in the source and target domains, respectively, N^s and N^t denote the number of training samples in the source and target domains, respectively. During training, the object detection network is optimized by minimizing the following loss function,

$$L_{YOLOv5} = L_{class} + L_{obj} + L_{loc} \tag{1}$$

where L_{YOLOv5} denote the total loss function of YOLOv5 [23]. L_{class}, L_{obj}, L_{loc} denote the classification loss, confidence loss and localization loss, respectively.

In addition, in order to make W_0, W_1, W_2 have the ability to distinguish features from the source domain or the target domain, the following loss function is minimized to optimize,

$$L_{id}(W_o) = \sum_{o=0}^{2} CE\left(W_o\left(E_o\left(\mathbf{x}_i^s\right)\right), y_0\right) + CE\left(W_o\left(E_o\left(\mathbf{x}_i^t\right)\right), y_1\right) \tag{2}$$

where $CE(\cdot)$ denotes the cross-entropy loss, E_o ($o = 0, 1, 2$) denote the features output by different network depths of E (the details can be seen in Figure 2), $y_0 = 0$ and $y_1 = 1$ denote the domain labels of the source and target domains, respectively.

Through model pre-training, the object detection network has a basic detection ability and W_0, W_1, W_2 can distinguish the source domain and target domain samples. However, when testing on unseen smoke datasets, the detection performance will drop sharply due to the domain shift between different datasets. In order to reduce the data distribution difference between the source domain and the target domain, this paper proposes Multilevel Feature Cooperative Alignment module.

3.2 Multilevel feature cooperative alignment (MFCA)

In order to mitigate the impact of inter-domain differences, this paper proposes a Multilevel Feature Aooperative domain Alignment module. The gap between the source and target domains is narrowed by adversarial learning between E and W_0, W_1, W_2. In theory, if the extracted features from source and target domain do not have

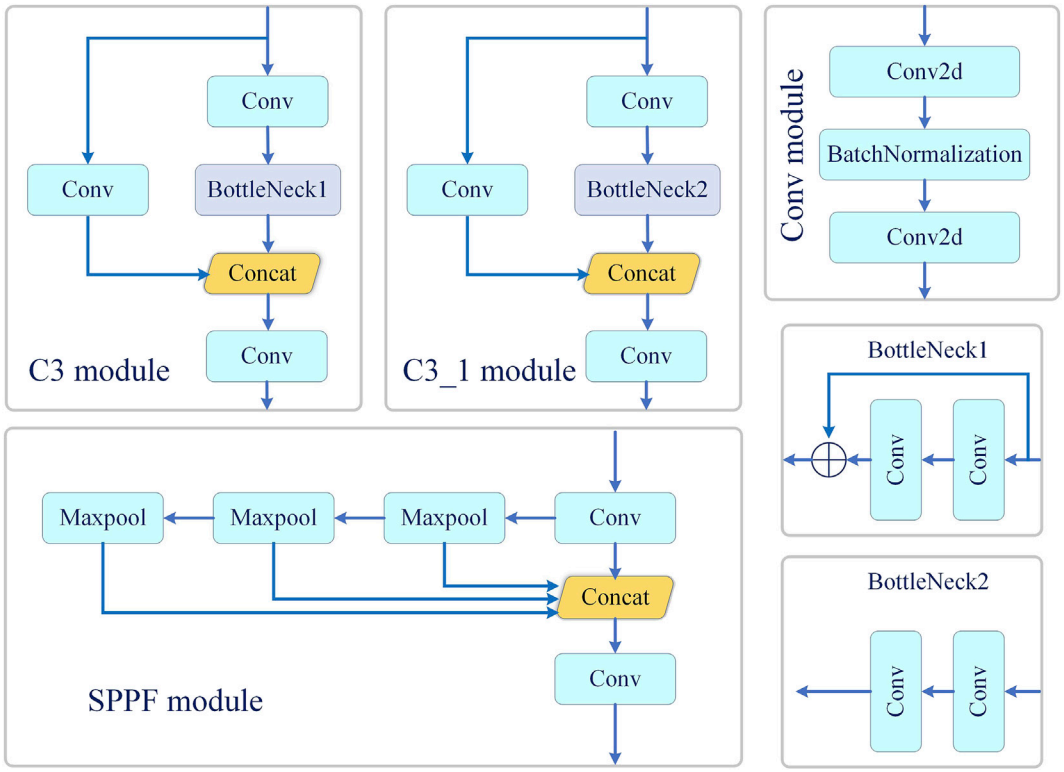


FIGURE 5
The details of the C3 module, SPPF module and other module in Figure 2.

differences, W_0, W_1, W_2 should not be able to distinguish the source domain and the target domain. By minimizing the following loss function to realize,

$$L_{adv}(E) = \sum_{o=0}^{2} CE\left(W_o\left(E_o\left(\mathbf{x}_i^s\right)\right), y_1\right) + CE\left(W_o\left(E_o\left(\mathbf{x}_i^t\right)\right), y_0\right) \quad (3)$$

At this time, the parameters of E (where E include E_0, E_1, E_2) are updated by fixing the parameters of W_0, W_1, W_2, and the source domain and target domain features are cross-constrained by domain labels. In this way, the extracted features are trained to adapt to the source domain and target domain, and the effect of domain alignment is achieved. The training process is shown in Figure 3. It is worth noting that, this paper not only uses the above way to mitigate the domain differences in the final features of Backbone, but also constrains the intermediate features of Backbone at the same time, and finally alleviates the impact of domain differences through the cooperative alignment of multi-level features.

3.3 Multilevel feature fusion (MFF)

In real scenes, smoke changes with time in different colors and shapes, which increases the difficulty of feature extraction. In order to improve the robustness of smoke features, this paper proposes to embed a Multilevel Feature Fusion module (MFF) in the Neck part of the detection network, the design of this module is shown in Figure 4.

Specifically, for the feature map $\boldsymbol{F}$ output by C3_1 module in Neck, it is divided into two branches. Follow CBAM [37], the first branch is enhanced by Channel Attention Module (CAM) and Spatial attention module (SAM), and the second branch is enhanced by 1×1 convolutional layer (Conv1×1) to adjust the feature dimensions and increase the non-linear mapping ability of the network. The deep features are further extracted, and then the correlation matrix $\boldsymbol{A}$ is calculated for the features of the two branches,

$$\boldsymbol{A} = Sigmoid\left(Conv\left[SAM\left(CAM\left(\boldsymbol{F}\right)\right); Conv\left(\boldsymbol{F}\right)\right]\right) \quad (4)$$

where, $Sigmoid\ (\cdot)$ represents the sigmoid activation function, $\boldsymbol{F}$ represents the output of C3_1 module in Neck, $[a; b]$ denotes concatenation of a and b, Conv denotes 1×1 convolution, SAM and CAM denote spatial attention module and channel attention module, respectively. The relation matrix $\boldsymbol{A}$ reflects the relationship between the corresponding positions of the feature maps obtained by the two branches, and the larger the value is, the more important it is. Finally, the feature maps of the two branches are fused by the following operations, and the fused features F_{Fusion} are used as the input of the later network layer.

$$F_{Fusion} = Conv\left(\left[\boldsymbol{A} \otimes F_{SAM}; (1 - \boldsymbol{A}) Conv\left(\boldsymbol{F}\right)\right]\right) \oplus \boldsymbol{F} \quad (5)$$

where, $F_{SAM} = SAM(CAM(\boldsymbol{F}))$ denotes the output of the spatial attention module, $\otimes$ denotes element-wise multiplication, and $\oplus$ denotes element-wise addition. It can be seen that the proposed fusion module adaptively adjusts the contribution of the two branches at the corresponding positions of $\boldsymbol{A}$ and 1-$\boldsymbol{A}$, so as to

TABLE 1 Comparison with other methods on RF → TS, RF → SF, SF → TS, and SF → RF. P and R denote Precision and Recall (%),respectively.

Methods	RF → TS			RF → SF			SF → TS			SF → RF		
	P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP
YOLOv3	1.79	17.4	12.7	3.69	71.3	78.2	1.33	15.9	8.72	2.85	85.7	59.0
YOLOv5s	21.4	18.6	14.3	84.4	74.6	80.3	16.9	9.77	9.20	64.5	64.8	58.4
Faster-RCNN	—	—	8.72	—	—	65.95	—	—	6.23	—	—	44.54
MCAF	33.6	23.9	21.6	86.9	82.8	88.5	30.6	17.6	14.3	67.4	66.7	66.0

TABLE 2 The running time of Faster-RCNN and MCAF in several setting.

Methods	RF → TS		SF → TS	
	Training time (/h)	Testing time (/s)	Training time (/h)	Testing time (/s)
Faster-RCNN	≈ 14.3	102.6	≈ 13.5	103.4
MCAF	≈ 1.5	7.2	≈ 1.3	7.1

achieve a better fusion effect, and finally improve the robustness of smoke features and enhance the representation ability of small target smoke.

3.4 Optimization

By considering all the loss functions jointly, the objective function in this paper is as follows,

$$L_{total}(E, W_o) = L_{YOLOv5} + L_{id}(W_o) + L_{adv}(E) \tag{6}$$

Firstly, the detection network and classifier are trained to have the basic smoke detection ability and the ability to distinguish the source domain and the target domain by L_{YOLOv5} and $L_{id}(W_o)$, respectively. Then, the adversarial learning strategy is used to alleviate the differences between the source domain and the target domain through $L_{adv}(E)$.

4 Experiments

In order to prove the effectiveness of the proposed method (MCAF), this chapter carries out a large number of experiments. Firstly, the data set used in the experiment is introduced, and then the performance of the proposed method is compared with that of classical object detection algorithms. Finally, the effectiveness and superiority of the proposed method are demonstrated by ablation experiments.

4.1 Datasets and evaluation protocol

The real scene dataset *True_smoke* (*TS*) used in this experiment contains a total of 4,128 images. Among them, 1,275 images were taken from real transmission lines, 2,853 images were taken from google search engine and State Key Laboratory of Fire Science of University of Science and Technology of China. The training set and test set were divided according to a ratio of 7:3, and LabelImg was used for annotation, and the annotation format was the same as that of the popular dataset PAS-CAL VOC [38], the annotation information was stored in the. *xml* file. In addition, the synthetic datasets *RFdataset* (*RF*) [10] and *SFdataset* (*SF*) [10] are also used for experiments. *RFdataset* (*RF*) is synthesized from real smoke and forest background and contains 12,620 images, where, 3,155 images are used for training and 6,310 images are used for testing. The *SFdataset* (*SF*) is synthesized from simulated smoke and forest background and contains 12,620 images, where, 3,155 images are used for training and 6,310 images are used for testing.

In this paper, Precision, Recall and mean average precision (mAP) are used as performance evaluation indicators. It is calculated as follows,

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{9}$$

where, *TP* is the number of samples correctly predicted as smoke, *FP* is the number of samples correctly predicted as smoke, *FN* is the number of samples correctly predicted as smoke, *N* is the total number of classes, and AP_i is the Average Precision of class *i*.

4.2 Implementation details

Considering that the actual application needs to deploy the algorithm to mobile devices, the YOLOv5s object detection network is used as the basic framework, and the detailed structure of each module is shown in Figure 5. In the training phase, the

TABLE 3 Ablation study. P and R denote Precision and Recall (%),respectively. MCAF represents the proposed method.

Methods	RF → TS			SF → TS		
	P	R	mAP	P	R	mAP
Baseline	21.4	18.6	14.3	16.9	9.77	9.2
Baseline+MFCA	33.1	20.2	17.2	18.1	16.8	11.2
Baseline+MFF	33.8	20.4	18.6	26.6	14.0	12.3
Baseline+MFCA+MFF (MCAF)	34.6	23.9	21.6	30.6	17.6	14.3

10.3389/fphy.2023.1136021

FIGURE 6
Detection result display. The left is the result of the proposed method, and the right is the result of baseline.

maximum of epochs is set to 100, in which the first 20 epochs are fore pretrain. Being similarly to YOLOv5, Mosaic, random cropping, horizontal flipping, etc., are used for data augmentation. the size of the input image is uniformly resized to 640 × 640×3 (for length, width and channel), and the feature maps output by the layer concatenated with Neck in backbone are used as the input of MFCA module (specifically, the outputs of the 4th, 6th and 9th layers of backbone are used as the input of MFCA module). The feature map sizes of the input three domain classifiers are 80 × 80×128, 40 × 40×256, and 20 × 20×512, respectively. The three classification networks have the same structure, consisting of global average pooling and fully connected layers. In addition, the MFF module is embedded behind the C3_1 module in Neck. For the training of Backbone, the SGD optimizer was used with the learning rate set to 0.01 and momentum to 0.3, and for the training of the three classification networks, the SGD optimizer was used with the learning rate set to 0.1 and momentum to 0.9. The batchsize for training and testing both set to 16. This experiment was performed on pytorch 1.13 [39], and all experiments were done on a Linux server for NVIDIA GEForce RTX3090Ti.

4.3 Comparison to other methods

At present, there is no public dataset for cross-domain smoke detection. In addition, existing works are trained under supervised conditions and cannot be directly compared. The comparison method in this paper uses more mature object detection methods. These methods include YOLOv3 [40], YOLOv5s [23], Faster-RCNN [27]. The comparative experimental Settings are RF → TS, SF → TS, RF → SF, SF → RF, (a → b represents a as the source domain and b as the

target domain cross-domain task), and the target domain category and location labels is unknown during training.

The experimental results are shown in Table 1. It can be seen that the mAP of the proposed method is much higher than that of classical object detection methods such as YOLOv3 and faster-RCNN. Such methods do not consider the inter-domain differences and thus perform poorly. Compared with YOLOv5s in the four experimental Settings, the mAP of the method in this paper is increased by 7.3%, 8.2%, 5.1%, 7.6% respectively, indicating that the method in this paper indeed enhances the ability of the model to extract smoke robust features.

Table 2 shows the comparison of training and testing time between the proposed method and Faster-RCNN. It can be seen that comparing with Faster-RCNN, the proposed method is more suitable for real-time smoke detection and more efficient.

4.4 Ablation study

This section discusses the ablation study. Firstly, the L_{YOLOv5}-guided optimized network is considered as Baseline. On the basis of Baseline, the Multilevel Feature Cooperative Alignment module and the Multilevel Feature Fusion module are gradually added, which proves that they are helpful to improve the performance, the results can be seen in Table 3. Ablation experiments were performed at RF → TS and SF → TS.

4.4.1 The effectiveness of multilevel feature cooperative alignment module (MFCA)

In order to alleviate the domain gap existing in the source domain and the target domain, a multilevel feature cooperative alignment module is proposed. By following the adversarial strategy of E and the domain classifier, removing the gap between the source and target domain. As shown in Table 3, Baseline on RF → TS, SF → TS Precision/Recall/mAP respectively is 21.4%/18.6%/14.3% and 16.9%/9.77%/9.2, When the Multilevel Feature Cooperative Alignment module is added, the performance is significantly improved, which proves the effectiveness of this module.

4.4.2 The effectiveness of multilevel feature fusion module (MFF)

In order to improve the feature representation ability of small target smoke, a multilevel feature fusion module is proposed. The feature fusion module is embedded in different depths of Neck in the detection network to enhance the features of different scales. As can be seen from Table 3, after adding the Multilevel Feature Fusion module on the basis of baseline, the Precision/Recall/mAP on RF → TS and SF → TS Raised to 33.8%/20.4%/18.6% and 26.6% 14.0%/12.3%. The validity of the module is proved.

As shown in Table 3, when the Multilevel Feature Cooperative Alignment module and the Multilevel Feature Fusion module are added to baseline at the same time, the overall performance is improved, indicating that the proposed method is effective.

In addition, Figure 6 shows the visualization of the detection results. Clearly, with the embedding of the proposed technique, the models become more powerful in terms of detection, confirming their significance.

5 Conclusion

Fire prevention is of great significance to the protection of human property safety, natural environment and industrial equipment. Smoke detection is helpful in the early warning of fire, and many researchers continue to improve the detection algorithm to meet the needs of this field. In order to adapt to smoke detection in real scenes, this paper proposes an unsupervised domain adaptive smoke detection algorithm based on multi-level feature fusion and cooperative alignment. On the one hand, the difference between the source domain and the target domain data is reduced by the cooperative alignment of features at different scales. On the other hand, by embedding fusion modules at different depths of Neck, the representation ability of features is enhanced. In this paper, the module structure, training method, loss function and network parameter setting of the proposed method are introduced in detail. The effectiveness of each module is proved by ablation experiments.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://smoke.ustc.edu.cn/datasets.htm.

Author contributions

FZ responsible for paper scheme design, experiment and paper writing. GW and YiM collectiong data. YW and YuM annotating data. GW and HP guide to do experiments and write papers. KW guide the paper scheme design and revision.

Funding

This work was supported by the Major scientific and technological projects of Yunnan Province Research on Key Technologies of ecological environment monitoring and intelligent management of natural resources in Yunnan (202202AD080010); Fundamental Research Fund of Science and Technology Department of Yunnan Province (202201AU070172).

Conflict of interest

Authors FZ, GW, YiM, YW, YuM, GW, and HP were employed by the Company Yunnan Power Grid Company Ltd.

The remaining author declares 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

References

1. Li H, Yan S, Yu Z, Tao D. Attribute-identity embedding and self-supervised learning for scalable person re-identification. *IEEE Trans Circuits Syst Video Technol* (2019) 30: 3472–85. doi:10.1109/tcsvt.2019.2952550

2. Li H, Chen Y, Tao D, Yu Z, Qi G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Trans Inf Forensics Security* (2020) 16:1480–94. doi:10.1109/tifs.2020.3036800

3. Li H, Dong N, Yu Z, Tao D, Qi G. Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. *IEEE Trans Circuits Syst Video Technol* (2021) 32:2814–30. doi:10.1109/tcsvt.2021.3099943

4. Li H, Xu K, Li J, Yu Z. Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification. *Knowledge-Based Syst* (2022) 251:109315. doi:10.1016/j.knosys.2022.109315

5. Lin X, Li J, Ma Z, Li H, Li S, Xu K, et al. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18-24 June 2022; New Orleans, LA, USA (2022). p. 20973–20982.

6. Wang Y, Qi G, Li S, Chai Y, Li H. Body part-level domain alignment for domain-adaptive person re-identification with transformer framework. *IEEE Trans Inf Forensics Security* (2022) 17:3321–34. doi:10.1109/tifs.2022.3207893

7. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

8. Berman D, Avidan S, Treibitz T. Non-local image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 27-30 June 2016; Las Vegas, NV, USA (2016). p. 1674–1682.

9. 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

10. Zhang Q, Lin G, Zhang Y, Xu G, Wang J. Wildland forest fire smoke detection based on faster r-cnn using synthetic smoke images. *Proced Eng* (2018) 211:441–6. doi:10.1016/j.proeng.2017.12.034

11. Li Y-J, Dai X, Ma C-Y, Liu Y-C, Chen K, Wu B, et al. Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18-24 June 2022; New Orleans, LA, USA (2022). p. 7581–7590.

12. Hsu H-K, Yao C-H, Tsai Y-H, Hung W-C, Tseng H-Y, Singh M, et al. Progressive domain adaptation for object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision; March 1-5, 2020; Snowmass Village, CO (2020). p. 749–757.

13. Inoue N, Furuta R, Yamasaki T, Aizawa K. Cross-domain weakly-supervised object detection through progressive domain adaptation. 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. 5001–5009.

14. Kim T, Jeong M, Kim S, Choi S, Kim C. Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 12456–12465.

15. Kim S, Choi J, Kim T, Kim C. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 27 October 2019 - 02 November 2019; Seoul, Korea (South) (2019). p. 6092–6101.

16. RoyChowdhury A, Chakrabarty P, Singh A, Jin S, Jiang H, Cao L, et al. Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 780–790.

17. Wu A, Liu R, Han Y, Zhu L, Yang Y. Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 10-17 October 2021; Montreal, QC, Canada (2021). p. 9342–9351.

18. Saito K, Ushiku Y, Harada T, Saenko K. Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 6956–6965.

19. Zhu X, Pang J, Yang C, Shi J, Lin D. Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 687–696.

20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* (2020) 63:139–44. doi:10.1145/3422622

21. Khan S, Muhammad K, Hussain T, Del Ser J, Cuzzolin F, Bhattacharyya S, et al. Deepsmoke: Deep learning model for smoke detection and segmentation in outdoor environments. *Expert Syst Appl* (2021) 182:115125. doi:10.1016/j.eswa.2021.115125

22. Liu H, Lei F, Tong C, Cui C, Wu L. Visual smoke detection based on ensemble deep cnns. *Displays* (2021) 69:102020. doi:10.1016/j.displa.2021.102020

23. ultralytics. yolov5 (2020). Available at: https://github.com/ultralytics/yolov5 (accessed on may 18, 2020).

24. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 23-28 June 2014; Columbus, OH, USA (2014). p. 580–587.

25. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans pattern Anal machine intelligence* (2015) 37:1904–16. doi:10.1109/tpami.2015.2389824

26. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision; December 7 - 13, 2015; Santiago, Chile (2015). p. 1440–1448.

27. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: C Cortes N Lawrence, editors. *Advances in neural information processing systems*. New York: Curran Associates, Inc (2015).

28. 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; 27-30 June 2016; Las Vegas, NV, USA (2016). p. 779–788.

29. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: Single shot multibox detector. In: European conference on computer vision; 8-16 October; Amsterdam, The Netherlands. Springer (2016). p. 21–37.

30. Chen Y, Li W, Sakaridis C, Dai D, Van Gool L. Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 18-23 June 2018; Salt Lake City, UT, USA (2018). p. 3339–3348.

31. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 22-29 October 2017; Venice, Italy (2017). p. 2223–2232.

32. Yin Z, Wan B, Yuan F, Xia X, Shi J. A deep normalization and convolutional neural network for image smoke detection. *Ieee Access* (2017) 5:18429–38. doi:10.1109/access.2017.2747399

33. Gu K, Xia Z, Qiao J, Lin W. Deep dual-channel neural network for image-based smoke detection. *IEEE Trans Multimedia* (2019) 22:311–23. doi:10.1109/tmm.2019.2929009

34. Zhao Y, Zhang H, Zhang X, Chen X. Fire smoke detection based on target-awareness and depthwise convolutions. *Multimedia Tools Appl* (2021) 80:27407–21. doi:10.1007/s11042-021-11037-1

35. He L, Gong X, Zhang S, Wang L, Li F. Efficient attention based deep fusion cnn for smoke detection in fog environment. *Neurocomputing* (2021) 434:224–38. doi:10.1016/j.neucom.2021.01.024

36. Zhan J, Hu Y, Zhou G, Wang Y, Cai W, Li L. A high-precision forest fire smoke detection approach based on argnet. *Comput Electron Agric* (2022) 196:106874. doi:10.1016/j.compag.2022.106874

37. Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); September 8-14, 2018; Munich, Germany (2018). p. 3–19.

38. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. *Int J Comput Vis* (2010) 88:303–38. doi:10.1007/s11263-009-0275-4

39. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in neural information processing systems*. New York: Curran Associates, Inc (2019).

40. Redmon J, Farhadi A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 03 February 2023
DOI 10.3389/fphy.2023.1136093

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Gang Hu,
Buffalo State College, United States
Cheng Yin,
Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, China
Shuai Zhou,
Kunming University of Science and Technology, China

*CORRESPONDENCE
Xiaodong Zhang,
doctorzxd@hospital.cqmu.edu.cn
Rui Xu,
xurui203389@hospital.cqmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 02 January 2023
ACCEPTED 23 January 2023
PUBLISHED 03 February 2023

CITATION
Qiu Y, Jiang L, Peng S, Zhu J, Zhang X and Xu R (2023), Procedural outcome following and Hemodynamic imaging analysis for anterior communicating artery wide-necked aneurysms by four different stents assisted coil embolization. *Front. Phys.* 11:1136093. doi: 10.3389/fphy.2023.1136093

Procedural outcome following and Hemodynamic imaging analysis for anterior communicating artery wide-necked aneurysms by four different stents assisted coil embolization

Yulong Qiu†, Li Jiang†, Shixin Peng, Ji Zhu, Xiaodong Zhang* and Rui Xu*

Department of Neurosurgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

Background: Anterior communicating artery (AcomA) aneurysm is the most common intracranial aneurysm (IA) and has the highest rupture rate. Previously, the preferred surgical treatment for intracranial aneurysms was microsurgery clipping (MC). With the gradual maturation of endovascular treatment (EVT), an increasing number of patients are inclined to treat IA with EVT. In recent years, an increasing number of scholars have suggested that the preferred treatment for wide-necked aneurysms is stent-assisted coiling (SAC). Currently, there are few studies on comparative analyses of the procedural results of SAC in AcomA aneurysms.

Methods: We retrospectively reviewed all consecutively treated patients who received SAC for AcomA aneurysms between 12 February 2013, and 20 January 2021. The primary procedural outcome was the occlusion rate evaluated with the Raymond–Roy occlusion classification (RROC) assessed on DSA at follow-up. Safety assessment included 1) ischemic complications (asymptomatic ischemia; intrastent thrombosis; coils falling off plug; arterial dissection); 2) bleeding complications (SAH; ICH); and 3) death. Univariate and multivariate logistic regression analyses were performed to determine patient baseline and aneurysm characteristics associated with total aneurysm occlusion at follow-up. Hemodynamic analysis was performed in one representative case each of the four stents, and six hemodynamic parameters were chosen, including wall shear stress (WSS), cavity blood flow velocity (CBFV), residual blood in the aneurysm (RBA), neck blood flow velocity (NBFV), blood flow inflow (BFI); and inflow concentration index (ICI).

Results: A total of 154 patients who underwent EVT *via* SAC were enrolled for comparative analysis of procedural outcomes. The median age was 55 years, and 56.49% (87) were female. At the first (6–10 months), second (12–15 months) and last (24–48 months) follow-up, complete aneurysm occlusion was observed in 94.8%, 94.8%and 94.2% of patients, respectively. There were no differences regarding the occlusion rates stratified by stent. Each stent showed a variable decrease in all hemodynamic parameters.

Conclusion: Hemodynamic parameters all decreased significantly after SAC with all four different stents, and the effect of laser-cut stents on the hemodynamic decline of aneurysms appeared to be more significant than that of woven stents. No

significant difference was observed in the follow-up RROC grade among the four stents.

KEYWORDS

AcomA aneurysms, wide-necked aneurysms, SAC, hemodynamic, imaging analysis

1 Introduction

Intracranial aneurysm (IA) is a common cerebrovascular disease, and the anterior communicating artery (AcomA) is the most common site [1], accounting for the highest mortality rate of ruptured aneurysms [2].

AcomA aneurysms are usually clipped, but an increasing number of studies have shown that endovascular treatment (EVT) is effective [3]. Stent-assisted coiling (SAC) is commonly used to treat wide-necked aneurysms, which are defined as aneurysms with neck ≥4 mm or dome-to-neck ratio (DNR) < 2 [4]. Previous researches believe that SAC has a similar complication rate and lower recurrence rate than coiling alone [5]. In our therapeutic centre, we used a total of four stents for SAC, which were Solitaire AB stents, Enterprise stents, LVIS(JR) stents and LEO + (baby) stents. In general, the treatment effect was acceptable, and the complication rate and recurrence rate were low.

The creation and rupture of aneurysms are associated with hemodynamic factors [6, 7]. Currently, there are relatively few studies that simultaneously investigate the effects of these four different stents on hemodynamics. In addition, no study on the simultaneous effect of these four stents on the occlusion rate has been reported.

In this study, we deeply investigated the effect of multiple factors, including patient baseline characteristics, anatomical aneurysm details, and hemodynamic changes before and after stent placement on aneurysm recurrence to guide the use of the stents in the subsequent treatments.

2 Materials and methods

2.1 Patient selection

We retrospectively analyzed the data of all consecutive patients with AcomA wide-necked aneurysms who were diagnosed and treated at the Neurosurgery of the First Affiliated Hospital of Chongqing Medical University between 12 February 2013, and 20 January 2021. All patients receiving SAC treatment with the four stents (Solitaire™, Enterprise™, LVIS(JR)™, LEO(Baby)™) were selected and analyzed.

The main inclusion criteria for all cases were 1) the diagnosis of an AcomA wide-necked aneurysm, regardless of rupture or not, 2) SAC in an elective setting, and 3) Solitaire™, Enterprise™, LVIS(JR)™ or LEO (baby)™ stents, regardless of previous treatments (EVT or MC). The main exclusion criterion was that the image display was unclear.

For all selected patients, baseline and procedural characteristics were collected from medical records and imaging studies. We selected only specific and measurable measures, for patient baseline, such as age, gender, etc., whereas height, weight, etc., were not recorded because of failure to measure precisely or variations in data. In terms of previous history, we selected arterial hypertension or not, history of previous treatment of aneurysms, other factors such as medical history which may cause difficulty in recording details due to patient forgetting or deliberate concealment, etc. so were not included as parameters.

2.2 Antiplatelet and anticoagulation therapy

Many reports suggest that irregular administration of clopidogrel and aspirin may result in high thromboembolic event rates [8]; therefore, we followed regular and scientific antiplatelet therapy before treatment. All patients without previous antiplatelet therapy received loading dose dual antiplatelet therapy primarily with aspirin (300 mg st) and clopidogrel (300 mg st) before SAC. If the standard dual antiplatelet dose (aspirin 100 mg qd and clopidogrel 75 mg qd) had been taken for ≥3 days, the same standard dual antiplatelet dose was given once before treatment. At the beginning of the intervention, weight-adaptive heparin was administered intra-arterially, 2/3 mg/kg for the first time. If the procedure was still not finished after 1 h, half of the first dose was injected as a bolus, and the dose was tapered sequentially. After surgery, some patients received intravenous micropumps of tirofiban for several hours, overlapping with dual antiplatelet therapy for 4 h before deactivation. For the remaining patients, coagulation profiles were reviewed at 4 h postoperatively, and low molecular weight heparin (LMWH) was administered subcutaneously if no significant abnormalities were observed, bridging dual antiplatelet therapy until the second postoperative day. All patients after SAC took aspirin for at least 3 months and clopidogrel for at least 6 weeks every day. The dosage was adjusted according to the thromboelastography (TEG) and clopidogrel genotype (CYP2C19 enzyme). Genetic polymorphisms of the CYP2C19 enzyme result in different efficacy of clopidogrel, which is usually divided into four phenotypes clinically according to the different ability to metabolize clopidogrel: ultra fast metabolism (UM), fast metabolism (FM), medium metabolism (MM) and slow metabolism (SM). When a patient's clopidogrel genotype was expressed as UM or FM, we gave clopidogrel 75 mg qd. When clopidogrel genotype was expressed as MM, we gave clopidogrel 100 mg qd. When clopidogrel genotype was expressed as SM, we replaced clopidogrel with ticagrelor at an oral dose of 90 mg qd.

2.3 SAC treatment

EVT is an effective treatment modality for AcomA aneurysms [9], and previous studies have shown that SAC may improve the results of embolization by allowing more complete initial coiling. The success rate of SAC is higher than that of coiling alone, especially in the treatment of wide-necked aneurysms [1, 10]. In our cases, all operations were performed under general anesthesia using a SIEMENS™ biplane angiography system. Cerebral vessel access was usually established using a 6F Chaperon Guiding Catheter System. Subsequently, a 3-D rotational angiography was performed to plan the

TABLE 1 Comparison of seven technical parameters among four stents.

Technical parameter	Solitaire AB stent	Enterprise stent	LVIS(JR) stent	LEO+(Baby) stent
Structure design	Nitinol laser-cut stent with a closed-cell design and fully open structure on one side	Nitinol laser-cut stent with a closed-cell design	Nitinol woven stent with a closed-cell design	Nitinol woven stent with a closed-cell design
Supporting method	self-expanding	self-expanding	self-expanding	self-expanding
Stent diameter -length,mm	3.0*–20/30	4.5–14/22/28/37	2.5–13/17/23/34	2.5–12/18/25
	4.0*–15/20/30/40		3.5–17/18/22/2328/33	3.5–18/25/30/35/50
	5.0*–20/30		4.0–12/17/22/28/31	4.5–15/20/25/30/40/50/75
	6.0*–20/30		4.5–18/23/32	5.5–50/60/75
			5.5–30/33	
Diameter of the parent artery,mm	1.5–6.5	2.0–4.0	2.0–5.5	1.0–6.5
Stent hole size,mm	2–3*	1.5	1.0	0.9
Metal coverage,%	5–7*	3.7–6.1	16–19	16–19
Recyclable	100% release	90% release	80% release	90% release

Legend: * Solitaire AB, stent is superimposable since its fully open structure on one side, which in turn leads to its adjustable diameter, hole size and metal coverage.

operation. In all cases, we used the jailing technique, which could reduce navigation time and result in a lower incidence of thromboembolic events [8]: 1) a microcatheter was used to navigate to the aneurysm and fill in coils; 2) another microcatheter was used to navigate to the far end of the parent artery; and 3) the stent was partially released while coiling the IA was continued; and 4) the stent was completely released when the IA was completely coiled.

2.4 Stent characteristics

The use of each stent depends on the diameter of the parent artery, the condition of vascular tortuosity and the specific parameters of the IA, such as the location and DNR. Each stent has its specific diameter-length ratio, and the choice of stent diameter is determined according to the diameter of the parent artery, the length of the stent is determined according to the width of the aneurysm neck, both ends of the aneurysm neck need to be covered with a sufficiently long stent, and at the same time the stent cannot be too long to block branch arteries.

Solitaire™ is a self-expanding nitinol laser-cut stent with a closed-cell design and fully open structure on one side, combining the advantages of open-cell and closed-cell designs. It can be completely recovered twice, and the stent hole is adjustable. The advantages of this stent are mainly a higher success rate for SAC of small-size aneurysms and a lower risk of postoperative thrombosis [11].

Enterprise™ is also a self-expanding nitinol laser-cut stent with a closed-cell design. The conveying system provides excellent navigation and positioning, making it easier to transport and deploy, with only a 1.3% inaccurate deployment rate. Stents are deployed on the parent artery 5 mm above both sides of the aneurysm neck, with more intensive filling of the aneurysm sac and neck [8]. Our case contained a large number of ruptured aneurysms, and previous studies have shown that the Enterprise stent is safe and effective in SAC of ruptured wide-necked aneurysms [8].

LVIS™ is a self-expanding nitinol woven stent system with a closed-cell design. It has higher metal coverage, which may cause a greater reduction in velocity at the neck plane, and higher packing density, which may cause a greater reduction in velocity and WSS at the aneurysm. The greater hemodynamic alterations may cause lower recanalization in medium-sized aneurysms [12].

LEO™ is a self-expanding nitinol woven stent system with a closed-cell design. With two radio opaque standards along its full length, delivery systems that allow easier navigation and precise placement and potential for stent repositioning make it very useful for the treatment of complex cerebral aneurysms [13], in addition to lower perioperative morbidity [14]. Some scholars even believe that SAC with LEO stents may be considered the first-line treatment for wide-necked aneurysms [15].

A comparison of the technical parameters of the individual stents is detailed in Table 1.

2.5 Hemodynamic parameters

Hemodynamics are as important as morphology in the development of aneurysms [7]. Previous studies have confirmed that 53% of patients with AcomA aneurysms have variant anatomy of the vessels surrounding the AcomA [2], and the presence of a hypoplastic A1 segment is the only parameter independently associated with the presence of AcomA aneurysms in addition to aneurysm size [16]. Research has shown that the growth of aneurysms may be associated with high wall shear stress (WSS) [17], and IAs with large maximum widths and small neck diameters, which have a greater range of low WSS areas, may be prone to rupture [18]. As aneurysms are studied more intensively, the influence of hemodynamics should not be ignored.

In this study, four patients with four different stents were randomly selected, focusing on the following six parameters: 1) wall shear stress (WSS); 2) cavity blood flow velocity (CBFV); 3) residual blood in the aneurysm (RBA); 4) neck blood flow velocity

(NBFV); 5) blood flow inflow (BFI); and 6) inflow concentration index (ICI).

2.6 Procedural outcomes

Aneurysm details were analyzed, including the size of the aneurysms, the DNR, ruptured or not, the number of coils filled and the diameters of the parent arteries. The angiographic results were evaluated after intervention and angiography by DSA using the Raymond–Roy Occlusion Classification (RROC) [19], which defined Class I as complete obliteration, Class II as residual neck, and Class III as residual aneurysm. RROC I was defined as the primary angiographic outcome endpoint, and all occlusion rates were compared by the stent used.

The above factors in all follow-up patients were analyzed for significant associations with complete aneurysm occlusion. We recommend that the patients be reviewed for the first time 6 months after surgery. If negative (Class I), patients should be reviewed again 1 year later. If still negative, they should be reviewed 2–3 years later. SAC was performed whenever the review was positive (Class II or Class III). The reason for this is that after aneurysm clipping, 25% of Raymond II patients will have gradually increasing aneurysms; if Raymond III, 75% of aneurysms will increase [20]. Although no relevant reports have been found after SAC, overall, more than 20% of aneurysms recurred after EVT [21].

Perioperative complications that need attention mostly include 1) ischemic complications (asymptomatic ischemia; intrastent thrombosis; coils falling off plug; arterial dissection); 2) bleeding complications (SAH; ICH); and 3) death.

2.7 Statistical analysis

All data were analyzed using statistical methods, standard descriptive statistics were used for all data endpoints, the mean and median were used to represent the degree of centralization, and the standard deviation (SD) was used to represent the distribution. Categorical variables were compared using the chi-square test or Fisher's exact test. Continuous variables were assessed with the Mann–Whitney *U* test (non-normally distributed data). Univariate and multivariate logistic regression analyses were used to identify variables associated with total aneurysm occlusion. The odds ratio (OR) and adjusted OR (AOR) had a 95% confidence interval. $p < 0.05$ was considered statistically significant. Regarding the hemodynamics section, we provided the original data mentioned above and technical support provided by ArteryFLOW Science and Technology Co., Ltd.

3 Results

3.1 Study population

A total of 154 patients were selected; 82.47% (127) of patients were found due to subarachnoid hemorrhage, and 17.53% (27) of patients were found accidentally due to health checkup or radiological examinations owing to symptoms such as dizziness or headache. The median age was 55 years, and 56.46% (87) were female. The most common risk factor associated with IA was arterial hypertension, with 51.95% (80) detected. The mean dome diameter, the mean neck diameter and the mean DNR was 5.4 mm, 3.6 mm, and 1.5, respectively. As the AcomA is the most difficult to observe by angiography among the arteries of the circle of Willis, the mean diameter of the parent artery can only be estimated roughly, which was approximately 1.62 mm. The median number of implanted coils was 3. Table 2 shows an overview of patient characteristics and anatomical aneurysm details stratified by stent models.

3.2 Complications

Perioperative ischemic complications (11.69%, 18) were more common than hemorrhagic complications (5.19%, 8). There were no deaths in our cases during hospitalization and follow-up.In total, 6.49% (10) developed asymptomatic ischemia; the incidence was 9.09% 4) for the Solitaire stent, 2.86% 1) for the Enterprise stents, 5.88% 2) for the LVIS stents, and 7.32% 3) for the LEO stents (Table 3).

3.3 Procedural outcome

According to the Raymond-Roy Classification, 94.81% (146) of patients had total occlusion (RROC Ⅰ) at the first follow-up. The incidence of RROC I was 86.36% (38) after Solitaire stenting, 97.14% (34) after Enterprise, 100% (34) after LVIS, and 97.56% (40) after LEO. At the second follow-up, 94.81% (146) of patients had total occlusion (RROC Ⅰ). The incidence of RROC I was 97.73% (43) after Solitaire stenting, 94.29% (33) after Enterprise, 94.12% (32) after LVIS, and 92.68% (38) after LEO. At the third follow-up, 94.16% (145) of patients observed total occlusion (RROC Ⅰ). The incidence of RROC I was 93.18% (41) after Solitaire stenting, 97.14% (34) after Enterprise stenting, 94.12% (32) after LVIS stenting, and 92.68% (38) after LEO stenting (Figure 1).

3.4 Logistic regression analyses

As shown in Table 4, most data for patients using different stents did not show significant differences. Univariate logistic regression was then used to analyze the influence of different factors on the occlusion rate at three follow-up reviews. The results showed that the influence of stents on the occlusion rate was not statistically significant; however, increasing dome size and increasing DNR were factors related to aneurysm occlusion at all three follow-ups; for details, see Table 5. Multivariable logistic analysis confirmed increasing DNR (adjusted odds ratio (aOR), 0.020; 95% CI, 0.001–0.583; $p = 0.023$) as an independent factor associated with complete aneurysm occlusion at the third follow-up.

3.5 Hemodynamic alterations

As no significant difference in aneurysm complete occlusion rates by different stents has been demonstrated above, we focused on hemodynamic aspects. Figure 2 demonstrates an angiogram before and after placement of four stents and a schematic representation of stents within blood vessels. In Picture A of Figure 3, we can see that at

TABLE 2 Patient baseline characteristics and anatomical aneurysm details.

Baseline characteristics	All patients	Solitaire™	Enterprise™	LVIS™	LEO™
Sum total	154	44	35	34	41
Age, median, years	55	56	56	55	56
Female	56.49% (87)	45.45% (20)	54.29% (19)	76.47% (26)	53.66% (22)
Arterial hypertension	51.95% (80)	43.18% (19)	42.86% (15)	52.94% (18)	68.29% (28)
Previous Treatment	4				
Coiled	2	1	1	-	-
Clipped	2	1	-	1	-
Aneurysm Signs					
Ruptured	82.47% (127)	77.27% (34)	77.14% (27)	91.18% (31)	85.37% (35)
Dome, mean, mm (±SD)	5.4 (2.21)	5.7 (2.52)	5.8 (2.64)	4.9 (1.53)	5.0 (1.83)
Neck, mean, mm (±SD)	3.6 (1.26)	3.6 (1.26)	3.9 (1.68)	3.5 (1.01)	3.4 (1.02)
DNR, mean (±SD)	1.5 (0.33)	1.6 (0.36)	1.5 (0.39)	1.4 (0.26)	1.4 (0.26)
Vessel diameter, Mean, mm (±SD)	1.62 (0.28)	1.64 (0.24)	1.62 (0.28)	1.64 (0.28)	1.57 (0.33)

SD: standard deviation.

TABLE 3 Perioperative complications.

Perioperative complications	All patients (n = 154)	Solitaire™ (n = 44)	Enterprise™ (n = 35)	LVIS™ (n = 34)	LEO™ (n = 41)
Ischemic complications					
Asymptomatic ischemia	10	4	1	2	3
Intrastent thrombosis	5	2	1	0	2
Coils falling off plug	2	0	1	0	1
Arterial dissection	1	1	0	0	0
Bleeding complications					
SAH	5	1	1	1	2
ICH	3	1	0	1	1
Death	0	0	0	0	0

SAH: subarachnoid hemorrhage; ICH: intracerebral hemorrhage.

the bifurcations and corners of the arteries, flow streamlines and WSS were elevated. High flow velocity territory (V > 0.1 m/s) was observed throughout the arterial vessel, including the aneurysmal body. In particular, from the section view, the flow velocity in the region of the dome was significantly lower than that elsewhere in the aneurysm before SAC, which may be because it was difficult for blood flow to reach the top due to the large dome diameter. After the aneurysm was coiled with a stent, it could be seen from the velocity magnitude contour at the cutting plane and isovagine surface (V > 0.1 m/s) that the blood flow in the aneurysm cavity was almost stagnant; at the same time, WSS was also significantly reduced. From the streamlines with color-coded velocity magnitude, we could see that there was no obvious change in the blood flow at stent placement; that is, it did not affect the blood flow in normal blood vessels. All stent models showed similar properties in Figure 3; in particular, Pictures A and D show slower flow velocities at the aneurysm dome site before SAC as a result of a larger DNR, whereas Picture C shows a non-significant decrease in aneurysm WSS after SAC.

As shown in Table 6, all indices had different decreasing trends in each stent. The decrease rate in LVIS was minimal for almost all parameters, and the Solitaire stent was optimal in four of all six indices. The decline rate of WSS was 43.74% compared with 83.27% for the LEO stent, 79.88% for the Enterprise stent and 79.76% for the Solitaire stent. The decline rate of the BFI was 79.07% compared with 92.65% for the Solitaire stent, 83.79% for the Enterprise stent and 82.39% for the LEO stent. The decline rate of the RBA was 93.10% compared with 95.36% for the Enterprise stent and 94.62% for the Solitaire stent. The decline rate of the CBFV was 68.93% compared with 86.96% for the Solitaire stent, 80.19% for the Enterprise stent and 79.17% for the LEO stent. The decline rate of the NBFV was 68.42% compared with 88.13% for the Solitaire stent, 75.93% for the Enterprise stent and 74.07% for the LEO stent. The decline rate of the ICI was 74.48% compared with 93.48% for the Solitaire stent, 84.41% for the Enterprise stent and 82.59% for the LEO stent.

3.6 Advantages of laser-cut stents in hemodynamics

The data of the present study suggest that the effect of laser-cut stents on the hemodynamic decline of aneurysms appears to be more significant than that of woven stents. The mean decline rate of WSS was 79.82% for laser-cut stents compared with 63.51% for woven

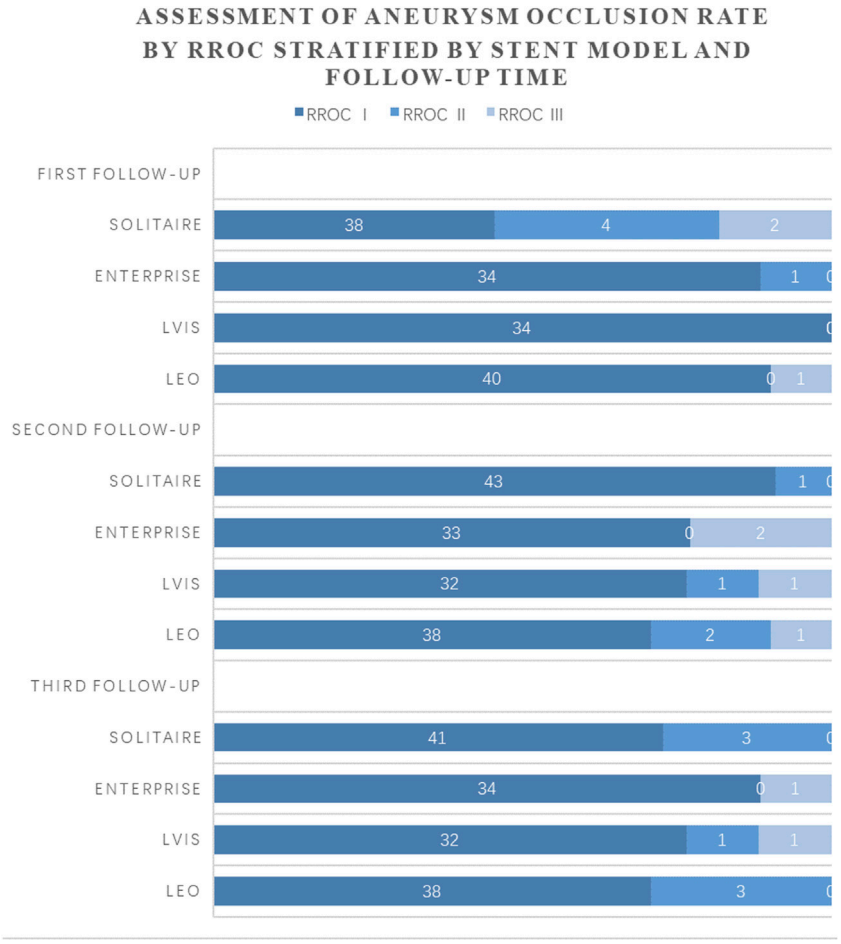


FIGURE 1
Comparison of occlusion rates at 3 follow-up visits in patients using four stents.Legend: Time of first follow-up: 6–10 months after SAC. Time of second follow-up: 12–15 months after SAC. Time of third follow-up: 24–48 months after SAC.

TABLE 4 Variability in patient data using each stent.

	P1 (Solitaire vs. Enterprise)	P2 (Solitaire vs. LVIS)	P3 (Solitaire vs. LEO)	P4 (Enterprise vs. LVIS)	P5 (Enterprise vs. LEO)	P6 (LVIS vs. LEO)
Age, years	0.929	0.263	0.644	0.317	0.587	0.078
Sex, female	0.442	0.005*	0.456	0.054	0.957	0.041*
Arterial hypertension	0.982	0.518	0.034*	0.555	0.048*	0.179
Dome size	0.851	0.093	0.139	0.075	0.113	0.770
Neck size	0.434	0.740	0.476	0.302	0.166	0.700
DNR	0.555	0.026*	0.057	0.150	0.276	0.572
Vessel diameter	0.681	0.940	0.246	0.775	0.494	0.348

Legend: * Indicating significance.

stents, of BFI was 88.22% compared with 80.73% for woven stents, of RBA was 94.99% compared with 87.66% for woven stents, of CBFV was 83.58% compared with 74.05% for woven stents, of NBFV was 82.03% compared with 71.25% for woven stents, and of ICI was 88.95% compared with 78.54% for woven stents.

4 Discussion

This study revealed several findings: 1) no significant difference was observed in follow-up RROC grade among the four stents; 2) ischemic complications were more frequent than hemorrhagic

TABLE 5 Univariate logistic regression analyses.

	First follow-up			Second follow-up			Third follow-up		
	OR	95% CI	*p* value	OR	95% CI	*p* value	OR	95% CI	*p* value
Age	0.955	0.884–1.032	0.248	0.978	0.904–1.058	0.584	1.011	0.937–1.091	0.780
Sex, female	0.769	0.177–3.338	0.726	0.769	0.177–3.338	0.726	0.077	0.018–1.227	0.077
Arterial hypertension	1.965	0.453–8.529	0.367	0.362	0.071–1.853	0.222	4.308	0.865–21.447	0.075
Dome size (increasing)	0.588	0.432–0.800	<0.001*	0.559	0.407–0.768	<0.001*	0.739	0.562–0.971	0.030*
Neck size (increasing)	0.699	0.432–1.131	0.144	0.708	0.437–1.147	0.160	0.915	0.548–1.526	0.732
DNR (increasing)	0.037	0.005–0.254	<0.001*	0.019	0.002–0.167	<0.001*	0.053	0.009–0.321	0.001*
Vessel diameter (increasing)	1.271	0.099–16.348	0.854	3.758	0.257–55.023	0.334	1.140	0.103–12.668	0.915
Solitaire (yes)	0.117	0.023–0.606	0.011*	2.922	0.349–24.476	0.323	0.788	0.188–3.302	0.745
Enterprise (yes)	2.125	0.252–17.884	0.488	0.876	0.169–4.546	0.875	2.450	0.296–20.294	0.406
LVIS (yes)	-	-	-	0.842	0.162–4.374	0.838	0.991	0.196–5.007	0.991
Leo (yes)	2.642	0.315–22.153	0.371	0.586	0.134–2.572	0.479	0.710	0.169–2.981	0.640

Legend: * Indicating significance. OR: odds ratio; CI: confidence interval.

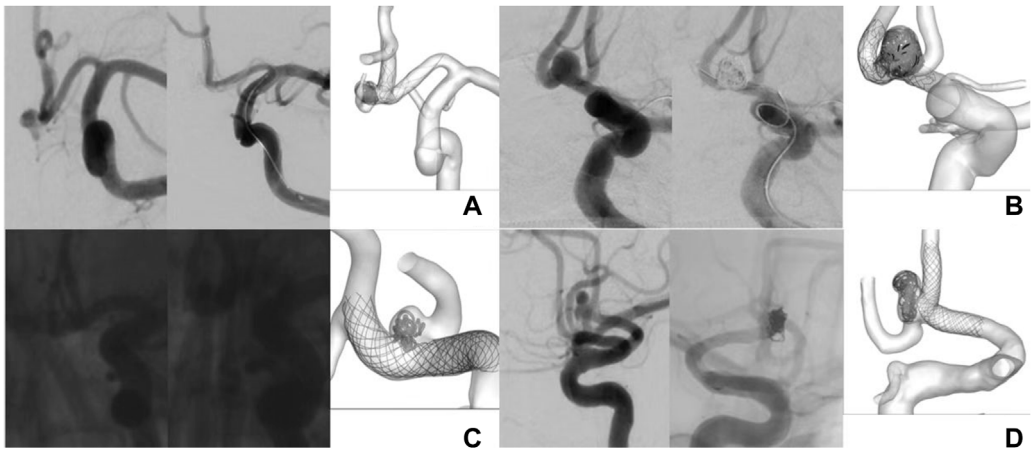


FIGURE 2
Angiogram before and after placement of four stents and schematic representation of stents within blood vessels. Picture **(A)** represents Solitaire stent; Picture **(B)** represents Enterprise stent; Picture **(C)** represents LVIS stent;Picture **(D)** represents LEO stent.

complications during our treatment; 3) hemodynamic parameters all decreased significantly after SAC with four different stents, and the effect of laser-cut stents on the hemodynamic decline of aneurysms appeared to be more significant than that of woven stents; and 4) increasing DNR was perhaps an independent factor associated with complete aneurysm occlusion.

It was previously thought that aneurysms first arise in high WSS areas, and subsequently, the growing areas change to areas of low WSS and eventually rupture there [17]. The latter is probably because low WSS (<1 Pa) is not sufficient for the self-healing process of the arterial wall, and the normal remodeling process of the arterial wall slows down, thereby leading to faster aneurysm growth and eventual rupture [22]. The WSS of the LEO stent decreased the most, which may be related to its characteristics as a woven stent with higher metal coverage or higher filling density. However, the LVIS stent, also as a woven stent, was slightly less effective in reducing WSS, suggesting that the fabrication process of the woven stent was perhaps not related to its function in reducing WSS; validation awaits further studies. Currently, the impact of WSS on aneurysm generation, growth, and rupture is recognized in two different ways by the academic community. These two schools of thought differ in the mechanisms leading to wall weakening. One view is that high WSS causes endothelial damage, which initiates chamber wall remodeling and potentially degeneration, leading to an imbalance between blood pressure and inner wall stress and resulting in local dilatation of the arterial wall. The resulting abnormal blood WSS fields lead to further aneurysm geometry development. An alternative view is that the arrest of flow locally against the wall within the dome at low WSS leads to endothelial dysfunction, with the aggregation and adhesion of platelets and leukocytes along the intimal surface, thereby causing intimal injury, inflammation, and subsequent degeneration of the canal wall. The aneurysmal wall will gradually become thinner, which may eventually lead to tearing of the tissue [23].

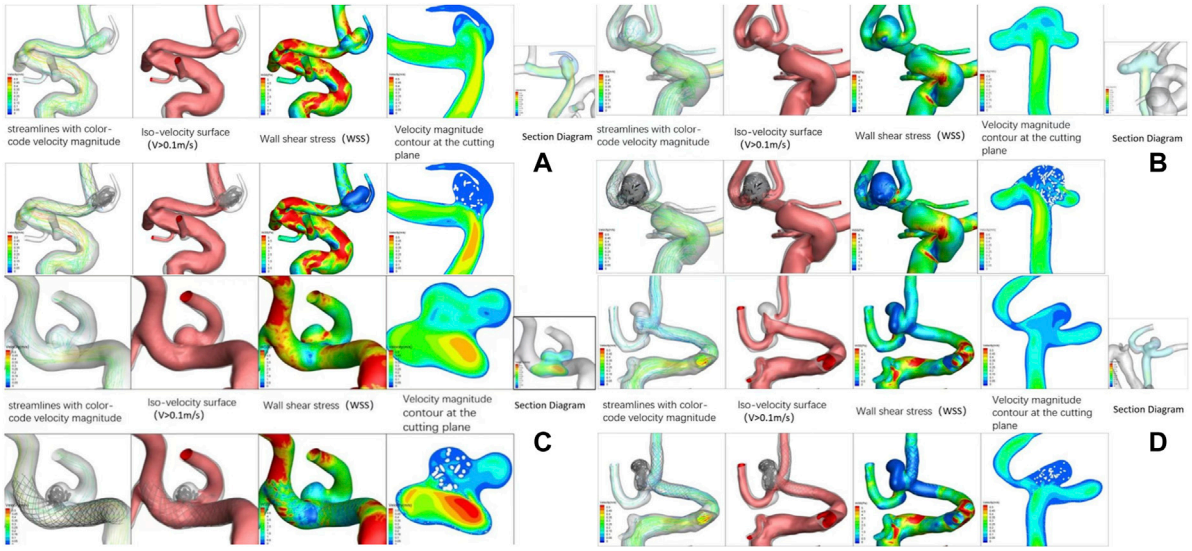


FIGURE 3
Hemodynamic patterns of four stents. Picture **(A)** represents Solitaire stent; Picture **(B)** represents Enterprise stent; Picture **(C)** represents LVIS stent; Picture **(D)** represents LEO stent.

TABLE 6 Changes of six hemodynamic indexes of the four stents before and after SAC.

	WSS (pa)	BFI (mL/s)	RBA (mm^3)	CBFV (m/s)	NBFV (m/s)	ICI
Solitaire						
Preoperative	1.981	1.130	52.0	0.161	0.160	1.534
Postoperative	0.401	0.083	2.8	0.021	0.019	0.100
Decline rate, %	79.76	92.65	94.62	86.96	88.13	93.48
Enterprise						
Preoperative	1.471	1.32	90.6	0.106	0.108	2.175
Postoperative	0.296	0.214	4.2	0.021	0.026	0.339
Decline rate, %	79.88	83.79	95.36	80.19	75.93	84.41
LVIS						
Preoperative	2.499	0.399	6.033	0.131	0.156	0.623
Postoperative	1.406	0.084	0.416	0.041	0.049	0.159
Decline rate, %	43.74	79.07	93.10	68.93	68.42	74.48
LEO						
Preoperative	0.538	0.477	13.5	0.048	0.054	1.407
Postoperative	0.090	0.084	2.4	0.010	0.014	0.245
Decline rate, %	83.27	82.39	82.22	79.17	74.07	82.59

WSS: wall shear stress; CBFV: cavity blood flow velocity; RBA: residual blood in the aneurysm; NBFV: neck blood flow velocity; BFI: blood flow inflow; ICI: inflow concentration index.

Stents were initially used in EVT to prevent the coils from slipping out but were found in clinical trials to straighten the parent artery and cause a corresponding change in the aneurysm sac, thereby altering aneurysm hemodynamics [24]. In sidewall aneurysms, the use of stents has previously been shown to block the inflow of flow and thereby reduce the corresponding hemodynamics [25], but in bifurcation aneurysms, it has been argued that stents may produce a greater inertial force resulting from narrowed inflow jet, which enhances flow into the aneurysms [26]. AcomA aneurysms are bifurcation aneurysms; however, in our study, the BFI (blood flow inflow) of each case was decreased, and the RBA (residual blood in the aneurysm) was decreased as well. Previous research suggests that aneurysm recurrence at >1 year after coiling is associated with higher intra-aneurysmal flow before and after coiling [27], which perhaps indicates the role of the four stents in improving the occlusion rate, and the performance of the Solitaire stent in reducing blood flow was especially excellent among the four stents. The most significant decline in RBA was in the case of LEO stents, perhaps related to the size of the stent hole and the density of coil embolization; thus far, no relevant studies have been reported.

Previous studies suggest that stents with lower porosity will cause a greater decrease in blood flow velocity [25], which was proven in our study in that the Solitaire™ stent had the largest decline rate in CBFV (cavity blood flow velocity) and NBFV (neck blood flow velocity)

among the four stents (Solitaire™ possessed the largest pore size and the lowest porosity of the four stents), and the decrease in blood flow velocity is crucial to prevent recanalization after EVT [25], which proved to be one of the advantages of this kind of stent.

ICI (inflow concentration index) refers to the degree of concentration of blood flow into the aneurysm. It is defined as the percentage of the flow rate of the parent artery which enters the aneurysm divided by the percentage of the aneurysm ostium area corresponding to positive inflow velocity [23]. Studies have found that ruptured aneurysms are more likely to have a concentrated inflow jet than unruptured aneurysms [23, 28]. All four stents were effective in reducing ICI and may be effective in reducing the incidence of aneurysm rupture during and after SAC. The ability of the Solitaire stent to reduce ICI is best, perhaps because of its greater stiffness as a laser-cut stent that is less pliable.

This paper is the first to study four types of stents (Solitaire™, Enterprise™, LVIS(JR)™, LEO+(Baby)™) simultaneously. Previous studies on the effects of Atlas ™, Enterprise ™, and LEO ™ stents on angiographic outcomes after SAC of wide-neck aneurysms showed a high incidence of total occlusion at long-term follow-up [29]. No significant difference was similarly observed among different stents, and the results of this research coincided with this finding. The four stents all showed varying degrees of reduction in terms of hemodynamics. In general, the reduction in hemodynamic factors by LVIS appeared to be minimal with all four stents; however, the final complete occlusion rate was not significantly different. The effect of laser-cut stents on the hemodynamic decline of aneurysms appears to be more significant than that of woven stents, which is probably caused by the different radial strengths of the stents due to the different fabrication processes. The present study demonstrated that increasing DNR (dome-neck ratio) is perhaps an independent factor associated with complete aneurysm occlusion. Other parameters related to aneurysm size and morphology, such as LWR (length-width ratio), size ratio (height-diameter of parent artery ratio), and HNR (height-neck ratio), and whether they correlate with aneurysm occlusion rates, require further study.

It was previously believed that it was unnecessary to treat small aneurysms found incidentally [30]. However, in this study, 45% (57) of the ruptured aneurysm domes were less than 5 mm. Advances in neuroradiological techniques have significantly improved the detection rate of small unruptured aneurysms, but whether to treat microaneurysms remains to be discussed.

Low-dose aspirin alone has previously been shown to reduce the risk of aneurysm rupture and is not associated with a risk of ICH compared with no therapy, but clopidogrel does have an increased risk of SAH and ICH [31]. Vigilance is necessary when administering dual antiplatelet therapy preoperatively and postoperatively, and TEG assays may be performed if necessary. In addition, previous studies suggested that antiplatelet drugs might change blood viscosity [32], which may affect the study of hemodynamics, which is not discussed in this study. Further studies are available later.

4.1 Limitations

The number of cases collected by different stents is different. Individual data values cannot be very accurate due to blurred images or personal capabilities [33]. Due to measurement factors, there may be some errors between aneurysm size and vessel diameter and the real situation. The stents are all closed-cell designs, with no involvement of open-cell stents. There were too few cases for hemodynamic studies, and the OSI (oscillatory shear index) was not studied.

5 Conclusions

Hemodynamic parameters all decreased significantly after SAC with four different stents, and the effect of laser-cut stents on the hemodynamic decline of aneurysms appeared to be more significant than that of woven stents. Although the reduction in hemodynamic factors by LVIS appeared to be minimal with all four stents, no significant difference was observed in follow-up RROC grade among the four stents. Ischemic complications were more frequent than hemorrhagic complications during our treatment, perhaps related to the operator's maneuvers and antiplatelet and anticoagulation therapy. Increasing DNR is perhaps an independent factor associated with complete aneurysm occlusion.

6 Future perspectives

Most recurrences occur within the first year after treatment; however, there is no universally agreed-upon timetable for imaging and clinical follow-up of treated aneurysms. Some scholars believe that longer follow-up should be considered for some types of high-risk aneurysms [34]. We sincerely hope that more multicenter studies on more stents with more long-term follow-up will be performed in the future.

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 human participants were reviewed and approved by Ethical Committee of the First Affiliated Hospital of Chongqing Medical University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

XZ and RX conceived the idea of the study; YQ and SP collected and analysed the data; LJ and JZ interpreted the results; YQ wrote the paper; all authors discussed the results.

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

References

1. Yarahmadi P, Kabiri A, Bavandipour A, Jabbour P, Yousefi O. Intra-procedural complications, success rate, and need for retreatment of endovascular treatments in anterior communicating artery aneurysms: A systematic review and meta-analysis. *Neurosurg Rev* (2022) 45(5):3157–70. doi:10.1007/s10143-022-01853-w

2. Luckrajh Jeshika S, Harrichandparsad R, Satyapal KS, Lazarus L. A clinical investigation of the anatomy of the proximal anterior cerebral artery and its association with anterior communicating artery aneurysm. *Translational Res Anat* (2022) 27:100200. doi:10.1016/J.TRIA.2022.100200

3. Yao L, Wu Q, Yuan B, Wen L, Yi R, Zhou X, et al. Correlation between vascular geometry changes and long-term outcomes after Enterprise stent deployment for intracranial aneurysms located on small arteries. *World Neurosurg* (2021) 153: e96–e104. doi:10.1016/J.WNEU.2021.06.038

4. Bsat S, Bsat A, Tamim H, Chanbour H, Alomari SO, Houshiemy MNE, et al. Safety of stent-assisted coiling for the treatment of wide-necked ruptured aneurysm: A systematic literature review and meta-analysis of prevalence. *Interv Neurora.: J Peritherapeutic Neurora. Surg Procedures Relat Neurosci.* (2020) 26:547–56. doi:10.1177/1591019920945059

5. Yang H, Sun Y, Jiang Y, Lv X, Zhao Y, Li Y, et al. Comparison of stent-assisted coiling vs coiling alone in 563 intracranial aneurysms: Safety and efficacy at a high-volume center. *Neurosurgery* (2015) 77:241–7. doi:10.1227/NEU.0000000000000765

6. Gao B-L, Hao WL, Ren CF, Wang JW, Liu JF. Greater hemodynamic stresses initiated the anterior communicating artery aneurysm on the vascular bifurcation apex. *J Clin Neurosci* (2022) 96:25–32. doi:10.1016/J.JOCN.2021.12.005

7. Xiang J, Natarajan SK, Tremmel M, Ma D, Mocco J, Hopkins LN, et al. Hemodynamic-morphologic discriminants for intracranial aneurysm rupture. *Stroke* (2011) 42:144–52. doi:10.1161/STROKEAHA.110.592923

8. Liu A, Peng T, Qian Z, Li Y, Jiang C, Wu Z, et al. Enterprise stent-assisted coiling for wide-necked intracranial aneurysms during ultra-early (48hours) subarachnoid hemorrhage: A single-center experience in 59 consecutive patients. *J Neurora.* (2015) 42:298–303. doi:10.1016/j.neurad.2014.11.005

9. Hur CW, Choi CH, Cha SH, Lee TH, Jeong HW, Lee JI. Eleven year's single center experience of endovascular treatment of anterior communicating artery aneurysms: Focused on digital subtraction angiography follow-up results. *J Korean Neurosurg Soc* (2015) 58(3):184–91. doi:10.3340/jkns.2015.58.3.184

10. Raymond J, Darsaut TE, Bing F, Makoyeva A, Kotowski M, Gevry G, et al. Stent-assisted coiling of bifurcation aneurysms may improve endovascular treatment: A critical evaluation in an experimental model. *AJNR Am J Neuroradiol* (2013) 34(3):570–6. doi:10.3174/ajnr.A3231

11. Zhang J, Wang D, Li X. Solitaire AB stent-assisted coiling embolization for the treatment of ruptured very small intracranial aneurysms. *Exp Ther Med* (2015) 10(6): 2239–44. doi:10.3892/etm.2015.2826

12. Li W, Wang Y, Zhang Y, Wang K, Zhang Y, Tian Z, et al. Efficacy of LVIS vs. Enterprise stent for endovascular treatment of medium-sized intracranial aneurysms: A hemodynamic comparison study. *Front Neurol* (2019) 10:522. doi:10.3389/fneur.2019.00522

13. Lv X, Li Y, Jiang C, Yang X, Wu Z. Potential advantages and limitations of the Leo stent in endovascular treatment of complex cerebral aneurysms. *Eur J Radiol* (2011) 79(2): 317–22. doi:10.1016/j.ejrad.2010.06.021

14. Luo J, Lv X, Jiang C, Wu Z. Preliminary use of the Leo stent in the endovascular treatment of wide-necked cerebral aneurysms. *World Neurosurg* (2010) 73(4):379–84. doi:10.1016/j.wneu.2010.01.019

15. Lubicz B, Kadou A, Morais R, Mine B. Leo stent for endovascular treatment of intracranial aneurysms: Very long-term results in 50 patients with 52 aneurysms and literature review. *Neurora.* (2017) 59(3):271–6. doi:10.1007/s00234-017-1805-3

16. Zhang J, Can A, Lai PMR, Mukundan S, Jr, Castro VM, Dligach D, et al. Vascular geometry associated with anterior communicating artery aneurysm formation. *World Neurosurg* (2021) 146:e1318–25. doi:10.1016/j.wneu.2020.11.160

17. Wang Y, Leng X, Zhou X, Li W, Siddiqui AH, Xiang J. Hemodynamics in a middle cerebral artery aneurysm before its growth and fatal rupture: Case study and review of the literature. *World Neurosurg* (2018) 119:e395–e402. doi:10.1016/j.wneu.2018.07.174

18. Wan H, Ge L, Huang L, Jiang Y, Leng X, Feng X, et al. Sidewall aneurysm geometry as a predictor of rupture risk due to associated abnormal hemodynamics. *Front Neurol* (2019) 10:841. doi:10.3389/fneur.2019.00841

19. Roy D, Milot G, Raymond J. Endovascular treatment of unruptured aneurysms. *Stroke* (2001) 32(9):1998–2004. doi:10.1161/hs0901.095600

20. Spiessberger A, Vogt DR, Fandino J, Marbacher S. Formation of intracranial de novo aneurysms and recurrence after neck clipping: A systematic review and meta-analysis. *J Neurosurg* (2019) 132(2):456–64. doi:10.3171/2018.10.jns181281

21. Ferns SP, Sprengers ME, van Rooij WJ, Rinkel GJ, van Rijn JC, Bipat S, et al. Coiling of intracranial aneurysms: A systematic review on initial occlusion and reopening and retreatment rates. *Stroke* (2009) 40(8):e523–9. doi:10.1161/STROKEAHA.109.553099

22. Farhan M, Didarul IM, Tarik AM. A study on the computational hemodynamic and mechanical parameters for understanding intracranial aneurysms of patients with hypertension and atrial fibrillation. *Inform Med Unlocked* (2022) 32:101031. doi:10.1016/J.IMU.2022.101031

23. Cebral JR, Mut F, Weir J, Putman C. Quantitative characterization of the hemodynamic environment in ruptured and unruptured brain aneurysms. *AJNR Am J Neuroradiol* (2011) 32(1):145–51. doi:10.3174/ajnr.A2419

24. Leng X, Wan H, Li G, Jiang Y, Huang L, Siddiqui AH, et al. Hemodynamic effects of intracranial aneurysms from stent-induced straightening of parent vessels by stent-assisted coiling embolization. *Interv Neuroradiol* (2021) 27(2):181–90. doi:10.1177/1591019921995334

25. Kono K, Shintani A, Terada T. Hemodynamic effects of stent struts versus straightening of vessels in stent-assisted coil embolization for sidewall cerebral aneurysms. *PLoS One* (2014) 9(9):e108033. doi:10.1371/journal.pone.0108033

26. Jeong W, Han MH, Rhee K. The hemodynamic alterations induced by the vascular angular deformation in stent-assisted coiling of bifurcation aneurysms. *Comput Biol Med* (2014) 53:1–8. doi:10.1016/j.compbiomed.2014.07.006

27. Damiano RJ, Tutino VM, Paliwal N, Patel TR, Waqas M, Levy EI, et al. Aneurysm characteristics, coil packing, and post-coiling hemodynamics affect long-term treatment outcome. *J Neurointerv Surg* (2020) 12(7):706–13. doi:10.1136/neurintsurg-2019-015422

28. Chung BJ, Doddasomayajula R, Mut F, Detmer F, Pritz MB, Hamzei-Sichani F, et al. Angioarchitectures and hemodynamic characteristics of posterior communicating artery aneurysms and their association with rupture status. *AJNR Am J Neuroradiol* (2017) 38(11):2111–8. doi:10.3174/ajnr.A5358

29. Strittmatter C, Meyer L, Broocks G, Alexandrou M, Politi M, Boutchakova M, et al. Procedural outcome following stent-assisted coiling for wide-necked aneurysms using three different stent models: A single-center experience. *J Clin Med* (2022) 11(12):3469. doi:10.3390/jcm11123469

30. Sato K, Yoshimoto Y. Risk profile of intracranial aneurysms: Rupture rate is not constant after formation. *Stroke* (2011) 42(12):3376–81. doi:10.1161/STROKEAHA.111.625871

31. García-Rodríguez LA, Gaist D, Morton J, Cookson C, González-Pérez A. Antithrombotic drugs and risk of hemorrhagic stroke in the general population. *Neurology* (2013) 81(6):566–74. doi:10.1212/WNL.0b013e31829e6ffa

32. Lee CH, Jung KH, Cho DJ, Jeong SK. Effect of warfarin versus aspirin on blood viscosity in cardioembolic stroke with atrial fibrillation: A prospective clinical trial. *BMC Neurol* (2019) 19(1):82. doi:10.1186/s12883-019-1315-5

33. 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

34. Pumar JM, Mosqueira A, Blanco-Ulla M, Vazquez-Herrero F. Recanalization and rupture of a brain aneurysm completely occluded with a LEO stent nine years ago. *Interdiscip Neurosurg* (2021) 25:101231. prepublish. doi:10.1016/J.INAT.2021.101231

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 06 February 2023
DOI 10.3389/fphy.2023.1124485

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Zhiqiang Zhang,
Hefei University of Technology, China
Xinghua Feng,
Southwest University of Science and Technology, China
Minghang Zhao,
Harbin Institute of Technology, Weihai, China

*CORRESPONDENCE
Yongfang Mao,
yfm@cqu.edu.cn

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 15 December 2022
ACCEPTED 10 January 2023
PUBLISHED 06 February 2023

CITATION
Liu Y, Chen X, Mao Y, Chai Y and Jiang Y (2023), Fault diagnosis of sensor pulse signals based on improved energy fluctuation index and VMD.
Front. Phys. 11:1124485.
doi: 10.3389/fphy.2023.1124485

Fault diagnosis of sensor pulse signals based on improved energy fluctuation index and VMD

Yuhu Liu, Xiaolong Chen, Yongfang Mao*, Yi Chai and Yutao Jiang

College of Automation, Chongqing University, Chongqing, China

Variational mode decomposition (VMD) has been widely applied in sensors. However, the mode number and balance parameter seriously limit VMD application. To solve this problem, this study proposes a novel method, which combines an improved energy fluctuation index (IEFI) and modified VMD (MVMD). In the proposed method, IEFI provided better performance to resist interference from random impulses by considering the periodicity of fault feature components. Consequently, it is applied to determine the initial center frequency for MVMD, which fixed the problem of the mode number. Moreover, a novel balance parameter search strategy, which can adaptively determine the optimal balance parameter, is combined with MVMD whose stop condition is replaced by kurtosis to extract the fault feature. Simulation results indicated that the proposed method does well in detecting the feature of a periodic impulse signal from the signal polluted by some interference impulses. Moreover, the bearing fault diagnosis results demonstrate that the proposed method can accurately detect bearing fault features. Furthermore, the method was validated with bearing fault data. The results showed that the method can accurately extract the fault characteristics of the impulse signal and achieve fault diagnosis.

KEYWORDS

fault diagnosis, impulse signal, bearing fault, improved energy fluctuation index, modified VMD

1 Introduction

Industrial equipment and systems have been increasingly moving toward larger, more complex, and integrated features, which lead to increased uncertainty in the system operation. To ensure the safe operation of equipment, extracting fault characteristics from signals collected by the sensors is necessary to achieve the purpose of fault diagnosis [1]. Sensors collect a large amount of image [2, 3] and data information [4–8], based on which many functions can be implemented.

Recent studies show the effectiveness of vibration signals in fault diagnosis [9]. Meanwhile, the fault response of the bearing and gearbox serves as an impact component in the vibration signal [10]. Unfortunately, the impulse response from the early fault is often submerged by noise from other running components and environments because the impulse response is too weak. Thus, an effective impulse signal detection method is necessary to evaluate the operating status of the rotation machine. Envelope analysis can effectively detect impulse signals but is ineffective in low signal-to-noise ratio (SNR) data. WT works well in heavy noisy signals but is seriously limited by basic functions [11]. EMD and EEMD can adaptively decompose complex signals into server modals but lack the rigorous mathematical theory. Fortunately, variational mode decomposition (VMD) can decompose low SNR signals into server modes under the number of suitable modes and the balance parameter [12]. Meanwhile, Wang et al. [13]

investigated the filter property of VMD by simulation signals and found that VMD can be implemented to detect impulse signals. Additionally, Wang et al. [14] applied VMD to detect impulse components in the signal from a rotor system. The results indicate that VMD works better than EMD and EEMC. Li et al. [15] analyzed the signal from a wind turbine by combining VMD and blind-source separation to detect the bearing crack fault. Li et al. [16] introduced VMD to calculate the central frequency and combined it with data-driven time–frequency analysis to diagnose the gear fault. Diagnosing faults by VMD provides advantages to identify different health conditions [17].

Based on the aforementioned description, VMD has been widely applied in the fault detection field. However, the mode number and balance parameter are determined based on the experience in the aforementioned articles. To solve this problem, many researchers paid attention to determining the mode number and balance parameter, and some results can be summarized as follows: first, research combined VMD with some intelligent search algorithms, such as grasshopper optimization algorithm, salp swarm algorithm, and particle swarm optimization [18–21]. By using intelligent search algorithms, the mode number and balance parameter can be determined adaptively and effectively. However, accepting the computational efficiency is difficult. Second, research studies put forward some other methods whose mode number is based on the fast Fourier transformation (FFT) spectrum of the decomposition result, such as independence-oriented VMD, adaptive VMD, and detrended fluctuation analysis VMD (DFA-VMD) [22–24]. These methods can adaptively select system parameters. However, some parameters must be determined artificially, and the over-decomposition phenomenon frequently occurs in these methods. Meanwhile, some researchers used iteration methods to search system parameters for VMD. Such methods include coarse-to-fine VMD and tentative VMD, which are often designed in two stages, to determine the target sub-mode and refine the sub-mode to enhance the impulse component. Finally, the initial center frequency-guided VMD (ICF-VMD) method is proposed in Refs. 25–Refs. 28. Compared with other adaptive VMD methods, ICF-VMD works well to extract bearing fault features and has better computational efficiency [29]. ICF-VMD is also designed in two stages: to determine the center frequency by the energy fluctuation variance and to refine the balance parameter to enhance the fault feature. However, the energy fluctuation variance is sensitive to the random impulse, and the balance parameter search process is limited in a narrow range. Two drawbacks may explain the failure of extracting the bearing fault feature.

To solve the aforementioned problems and improve the computational efficiency, this study proposes a novel method which combines an improved energy fluctuation index (IEFI) and modified VMD (MVMD). IEFI, a method based on the original energy fluctuation index and the subscript's variance of the energy whose value is greater than the mean, is used to determine the center frequency for MVMD. Consequently, the mode number can be fixed as one, and the balance parameter is the only parameter that needs to be determined. In this research study, a novel balance parameter search strategy from MVMD was used to extract the bearing fault feature. The initial balance parameter is determined based on the center frequency from the IEFI, which enhances the adaptability of the search strategy. The MVMD, whose stop condition is replaced by kurtosis, has good computational efficiency. In summary, IEFI ensures that the proposed method works well to process the signal, which includes some random impulses. The novel search strategy and MVMD ensure the computational efficiency of the proposed method. The effectiveness of the proposed method is examined by the simulation and experiment signals. The advantages of the proposed method are highlighted by comparing it with some existing methods.

The rest of the paper is organized as follows. Section 2 describes the proposed method. Section 3 organizes the results of the numerical experiment, case study, and comparison. Section 4 presents a concise summary.

2 The proposed method

This section introduces the basic theory about the IEFI and the MVMD to help in understanding the proposed method.

2.1 Modified VMD

VMD decomposes signals into a series of sub-modes through some Wiener filter banks. Its model is described as follows:

$$\begin{aligned} &\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \\ &s.t. \sum_{k=1}^{K} u_k(t) = f \end{aligned} \tag{1}$$

where u_k and ω_k denote the sub-mode and its center frequency, respectively. By introducing Lagrangian multipliers and penalty technology, Eq. 1 can be written as follows:

$$\begin{aligned} L(\{u_k\},\{\omega_k\},\lambda) = \alpha \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \\ \left\| f(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 + \langle \lambda(t), f(t) - \sum_{k=1}^{K} u_k(t) \rangle \end{aligned} \tag{2}$$

where α is the balance parameter, and $\lambda(t)$ means the Lagrangian multiplier parameter. Equation 2 can be solved through an alternate direction method of multiplier (ADMM) technology, and its process is described in.

Algorithm 1: ADMM for VMD

Initialize: $u_k, \omega_k, \lambda, n \leftarrow 1$

$$\textbf{Update } u_k\text{: } \hat{u}_k^{n+1}(\omega) \leftarrow \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_k^{n+1}(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^n)^2}, \tag{3}$$

$$\textbf{Update } \omega_k\text{: } \omega_k^{n+1} \leftarrow \frac{\int_0^\infty \omega |u_k^{n+1}(\omega)|^2 d\omega}{\int_0^\infty |u_k^{n+1}(\omega)|^2 d\omega}, \tag{4}$$

$$\textbf{Update } \lambda\text{: } \lambda^{n+1}(\omega) \leftarrow \lambda^n(\omega) + \tau \left(f(\omega) - \sum_k u_k^{n+1}(\omega) \right). \tag{5}$$

Convergence condition: $\sum_k \|\hat{u}_k^{n+1} - \hat{u}_k^n\|_2^2 / \|u_k^{n+1}\|_2^2 < \varepsilon$

According to, τ denotes the learning rate, which can be fixed as zero when VMD is applied to denoise the sub-components instead of recovering them. Understanding the mode number and the balance

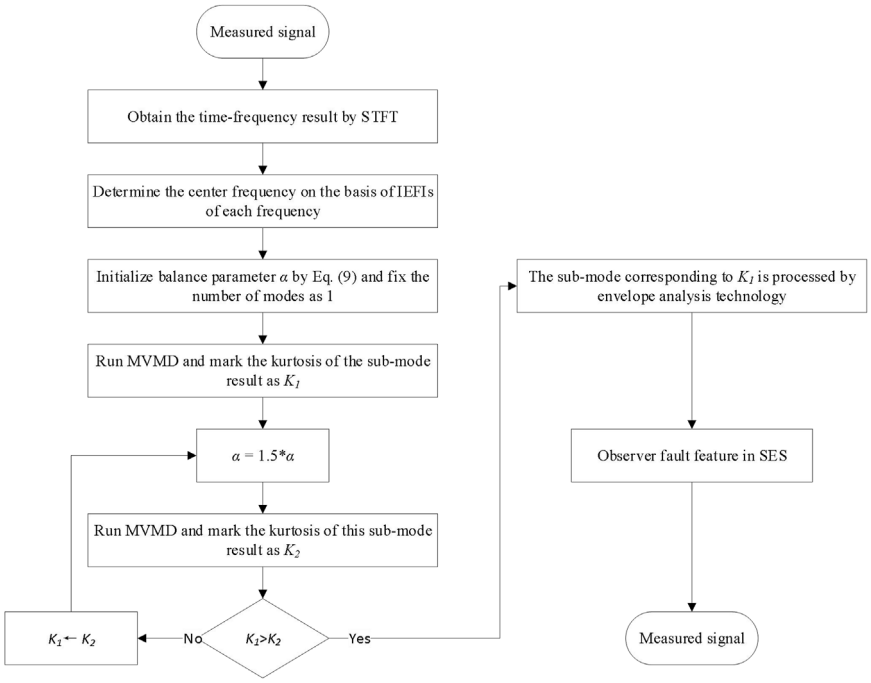


FIGURE 1
Flowchart of IEFI–MVMD

parameter is easy and important in VMD. According to Ref. 12, the mode number can be fixed as one with the help of the right center frequency. On the other hand, according to Ref. 30, this method is applied to detect bearing fault features. Thus, the mode number is set as one. This study applies envelope analysis to process the signal filtered by VMD. Consequently, VMD is assumed to be the filter in this study, and the learning rate τ should be 0 based on 11. Given that the purpose of VMD in this study is not to recover sub-components, its convergence condition can be modified to obtain a higher computational efficiency. Kurtosis is widely applied as index for diagnosing bearing fault, and it will be applied to construct a new convergence condition for VMD in this study. The new convergence condition is defined as follows:

$$\mathrm{kur}\left(u^n\right)/\mathrm{kur}\left(u^{n+1}\right) > \eta, \tag{6}$$

in which, $\mathrm{kur}(\cdot)$ denotes the kurtosis operator and η is set as 0.99, which can ensure that the kurtoses of the two adjacent generations have the same level. VMD, which is based on this convergence condition, is named MVMD in this study. By modifying the convergence condition as Eq. 6, MVMD not only has a good performance in extracting bearing fault features but also has higher computational efficiency, which is friendly with engineering applications.

2.2 Improved energy fluctuation index

Based on 11, Refs. 28, the number of modes can be set as one with the help of a correct center frequency. According to Ref. 28, the center frequency is determined based on the variance of energy fluctuations whose mathematical formula can be written as:

$$A\left(f_j\right) = \sqrt{\sum_{i=1}^{N}\left(TF\left(t_i, f_j\right), -\overline{TF\left(t_i, f_j\right)}\right)^2}, \tag{7}$$

where $TF(t, j)$ is the time–frequency analysis result. In this research, it is calculated by the short-time Fourier transform (STFT), which is shown as follows:

$$TF\left(t, f\right) = \int_{-\infty}^{+\infty} x(\tau) w(t-\tau) e^{-2j\pi f\tau} d\tau. \tag{8}$$

However, the variance of energy fluctuation is weak to resist the interferences from the random impulses and neglects the period property of the real fault response. To fill these gaps, an IEFI is proposed to determine the center frequency. The new index is defined as:

$$IEF\left(f_j\right) = A\left(f_j\right) \times \exp\left\{-\mathrm{var}\left(S_S\left(f_j\right) - S_F\left(f_j\right)\right)\right\}. \tag{9}$$

$S_S(f_j)$ corresponds to the subscripts of the elements from the second to the last, whereas $S_F(f_j)$ corresponds to the subscripts of the elements from the first to the last but one. For the periodic impulses, all of the elements in $[S_S(f_j) - S_F(f_j)]$ should be constant. Thus, their variance should be equal to zero. Therefore, the exponent term shown in Eq. 9 will be close to one for periodic impulses. However, for aperiodic impulses and noise, the distribution of the elements in $[S_S(f_j) - S_F(f_j)]$ is irregular. Thus, their variance is far from zero, which will weaken the exponent term shown in Eq. 9. Based on the aforementioned description, implementing IEFI to identify periodic impulses is more accurate than implementing raw energy fluctuations.

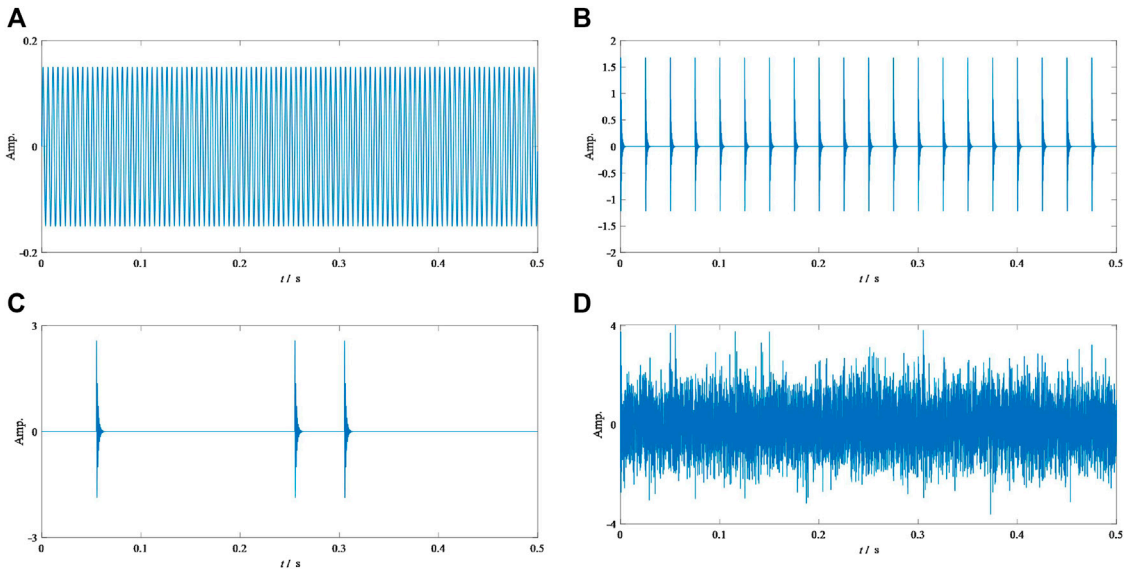


FIGURE 2
Simulation signal: **(A)** harmonic component y_1, **(B)** periodic impulse component y_2, **(C)** interface impulse component y_3, and **(D)** composite signal y.

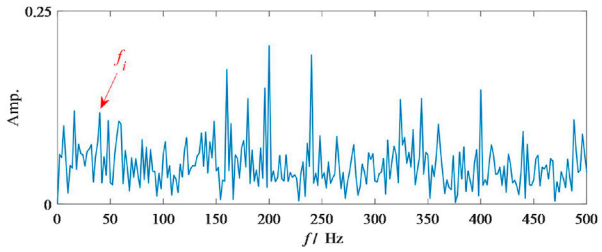


FIGURE 3
SES of the simulation signal.

2.3 The proposed method

The center frequency can be obtained from the IEFI, and the number of modes is set as one based on it. To run VMD successfully, the balance parameter should be determined first. Generally, the intelligent search algorithm and iterative search process are applied to solve this problem. However, accepting the computational efficiency of intelligent search algorithm is difficult. Thus, this study proposes a novel iterative search strategy to determine the balance parameter. The novel method is named IEFI–MVMD, which combines the improved energy fluctuation index and the modified VMD. The main steps of IEFI–MVMD are given as follows:

Step 1: The signal is processed by STFT with a window length of 512 and an overlap of 256.

Step 2: The IEFI is applied to evaluate the periodic impulse for each frequency. Additionally, the center frequency is the one with the largest IEFI.

Step 3: The balance parameter is initialized on the basis of

$$\alpha = \varphi / \min(0.5 - \omega_k, \omega_k)^2, \tag{10}$$

where φ is defined as five based on Eq. 3. From Eq. 3, one can easily understand that $\varphi = 5$ ensures the frequency response factor of the frequency boundary is not over 1/10.

Step 4: The raw signal is processed by using the MVMD method whose modes' number is fixed as one, and the kurtosis of the decomposition result is marked as K_1.

Step 5: The balance parameter is replaced by $\delta \times \alpha$, and δ is fixed as 1.5 in this study. The kurtosis of the new result is calculated and marked as K_2. If K_1 is less than K_2, then, Steps 4 and 5 are repeated until K_1 is larger than K_2.

Step 6: The final result (corresponding to K_1) is processed through envelope demodulation technology to obtain the squared envelope spectrum, which can clearly show the fault feature frequency.

To understand the IEFI–MVMD clearly, Figure 1 displays the corresponding flowchart.

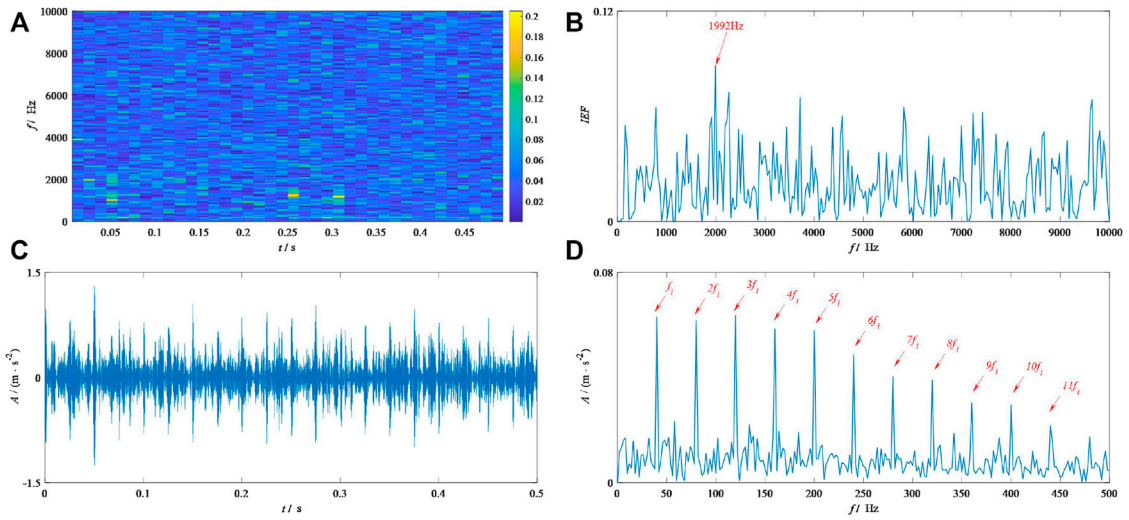


FIGURE 4
Results by the IEFI–MVMD of the simulation signal: **(A)** STFT, **(B)** IEFI, **(C)** TDW, and **(D)** SES.

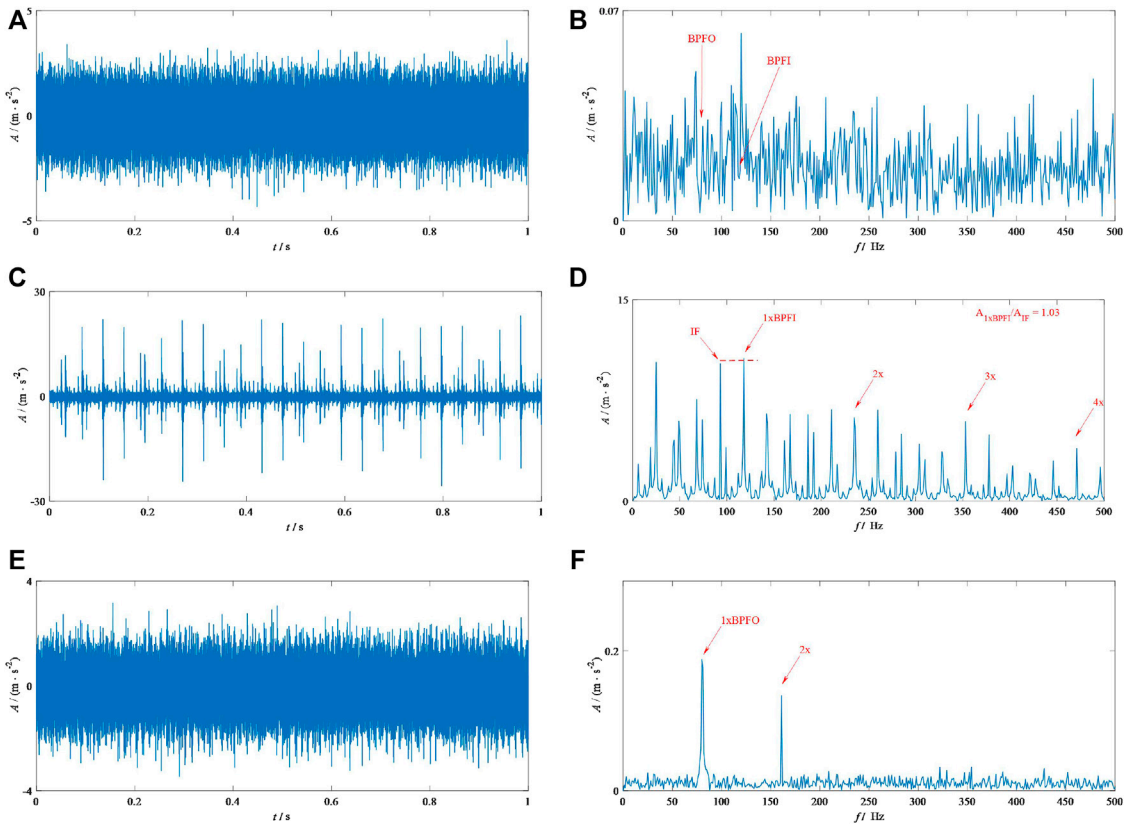


FIGURE 5
Signals from MFPT: **(A)** and **(B)** correspond to the TDW and SES of the healthy bearing, **(C)** and **(D)** correspond to the TDW and SES of the inner race fault bearing, and **(E)** and **(F)** correspond to the TDW and SES of the outer race fault bearing.

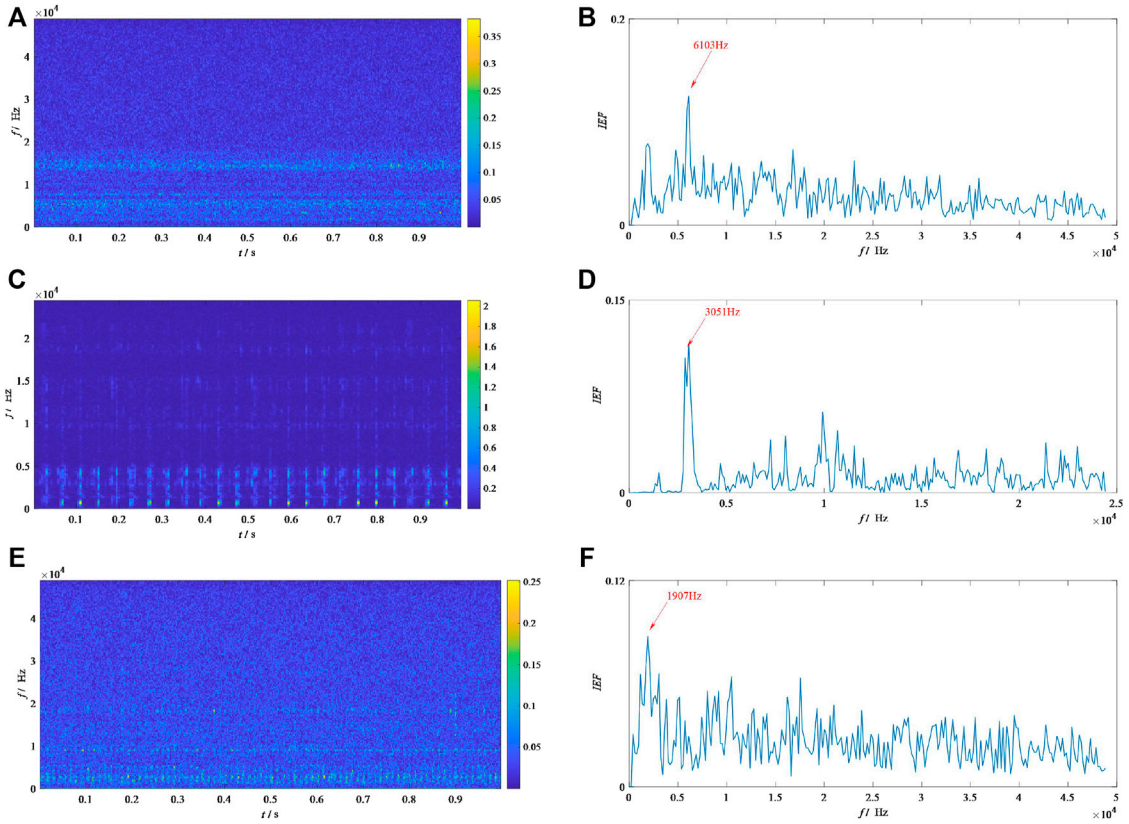


FIGURE 6
Results of IEFI for signals from MFPT: **(A)** and **(B)** correspond to the STFT and IEFI of the healthy bearing, **(C)** and **(D)** correspond to the STFT and IEFI of the inner race fault bearing, and **(E)** and **(F)** correspond to the STFT and IEFI of the outer race fault bearing.

3 Case study

To examine the effectiveness of the proposed method, this section introduces a simulation signal and two bearing fault signals. Meanwhile, the superiority of the proposed method is highlighted by comparing it with some existing methods.

3.1 Simulation

The simulation signal includes the harmonic component (y_1), the periodic impulse (y_2), the aperiodic impulse (y_3), and the Gaussian noise ($n(t)$). The simulation signal can be written as follows:

$$y(t) = y_1(t) + y_2(t) + y_3(t) + n(t), \tag{11}$$

$$y_1(t) = A \times \sin(2\pi \times 200t), \tag{12}$$

$$y_2(t) = 2 \times e^{0.1\times 2\pi \times 2000t} \sin\left(2\pi \times 2000t \times \sqrt{1-0.1^2}\right), \tag{13}$$

$$y_3(t) = 3 \times e^{0.1\times 2\pi \times 1200t} \sin\left(2\pi \times 1200t \times \sqrt{1-0.1^2}\right). \tag{14}$$

In the simulation signal, the frequency of the impulse signal is set at f_i = 40Hz. The sampling frequency is 20 kHz, and the length of the simulation signal is 10 k points. The density of Gaussian noise is 0.4. The simulation signal is illustrated in Figure 2. From Figure 2D, the periodic impulses can be seen as seriously polluted by the noise. Even in its squared envelope spectrum (SES) shown in Figure 3, observing the features of the periodic impulses is difficult.

Then, the proposed method is applied to analyze this signal. To begin with, the signal is analyzed by the IEFI, and Figure 4 shows the results. As shown in Figure 4A, three interference impulses occur in the simulation signal, which is consistent with the results shown in Figure 2C. Figure 4B shows the result above IEFI. Additionally, the frequency corresponding to the largest IEFI is 1,992 Hz, which is close to the design frequency in Eq. 13. Then, the balance parameter is determined by Eq. 9. Figure 4C is the time domain waveform (TDW) of the results. Compared to the raw TDW shown in Figure 2D, some periodic impulses are clearly shown in this figure. Importantly, the fundamental feature frequency and its harmonics are clearly displayed in its SES, as shown in Figure 4D. Consequently, our method succeeds in detecting the feature of the periodic impulses from the signal polluted by some interference impulses.

3.2 Case I

This section describes the implementation of the proposed method to analyze some signals from the bearing fault experiment. The signals used in this section come from the Society for Machinery Failure Prevention Technology (MFPT). According to description in MFPT, the tested bearing's faults include healthy conditions, outer race fault conditions, and inner race fault conditions. Figure 5 illustrates the TDW and the corresponding SES of these signals used in this section. From Figure 4, the amplitudes of the fault feature frequencies of the healthy bearing can be

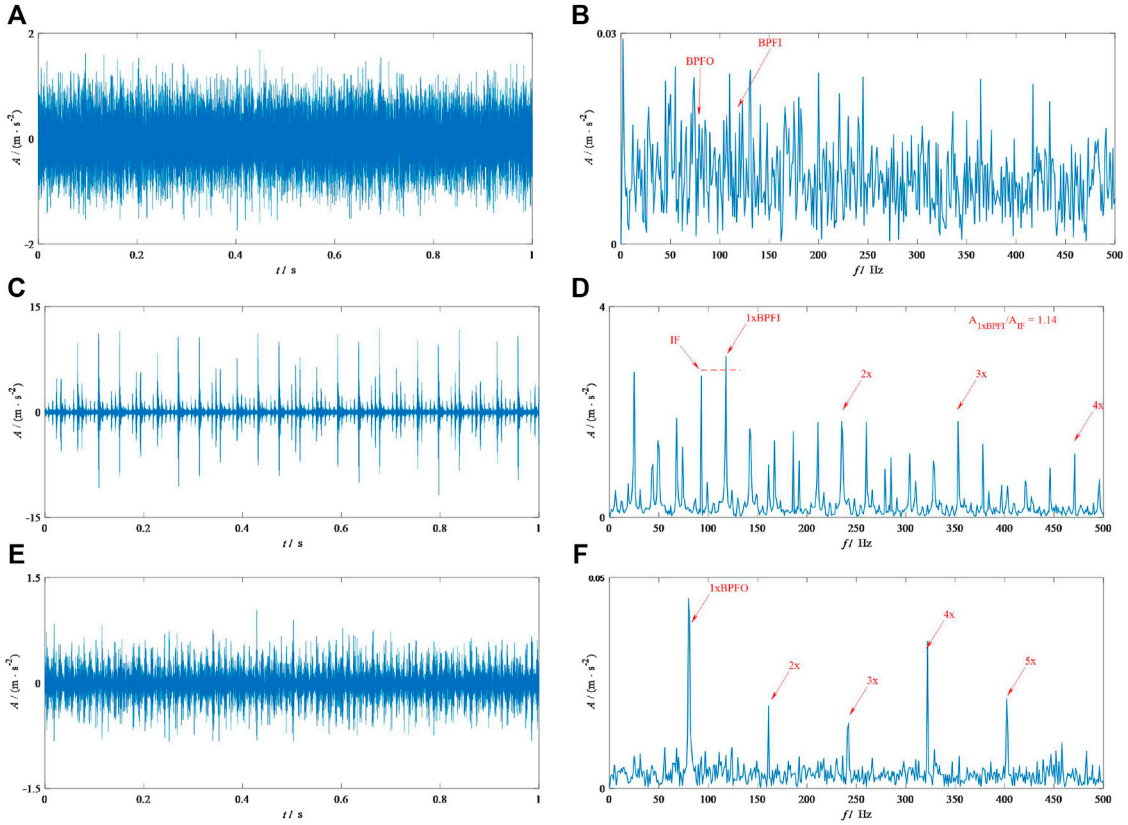


FIGURE 7
Final results of the IEFI–MVMD for signals from MFPT: **(A)** and **(B)** correspond to the TDW and SES of the healthy bearing, **(C)** and **(D)** correspond to the TDW and SES of the inner race fault bearing, and **(E)** and **(F)** correspond to the TDW and SES of the outer race fault bearing.

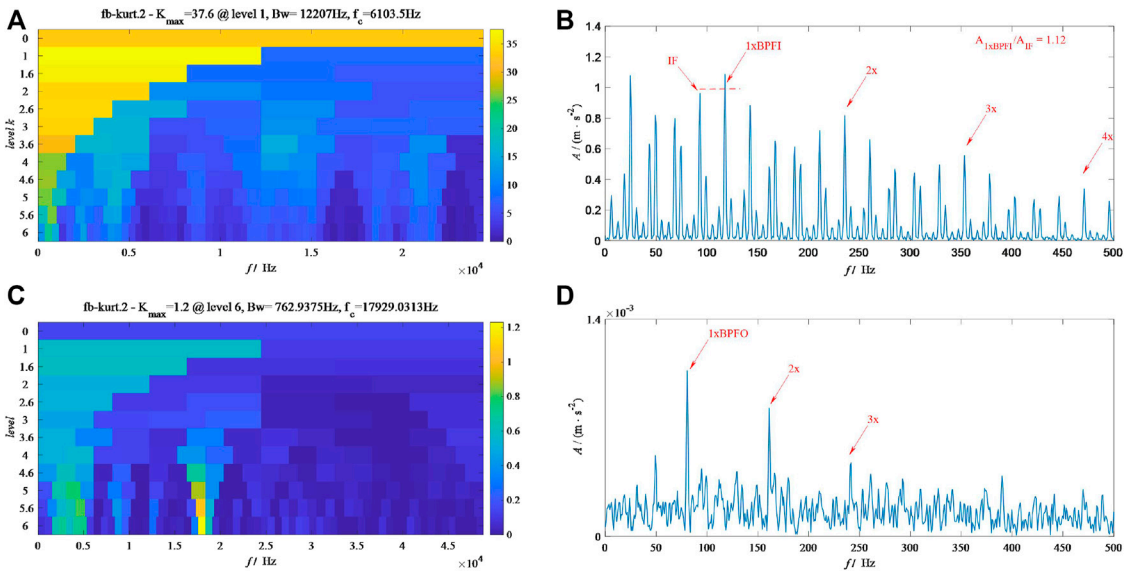


FIGURE 8
Results from FK for the MFPT signals: **(A)** and **(B)** correspond to the kurtogram and the corresponding SES of the inner race fault bearing; **(C)** and **(D)** correspond to the kurtogram and the corresponding SES of the outer race fault bearing.

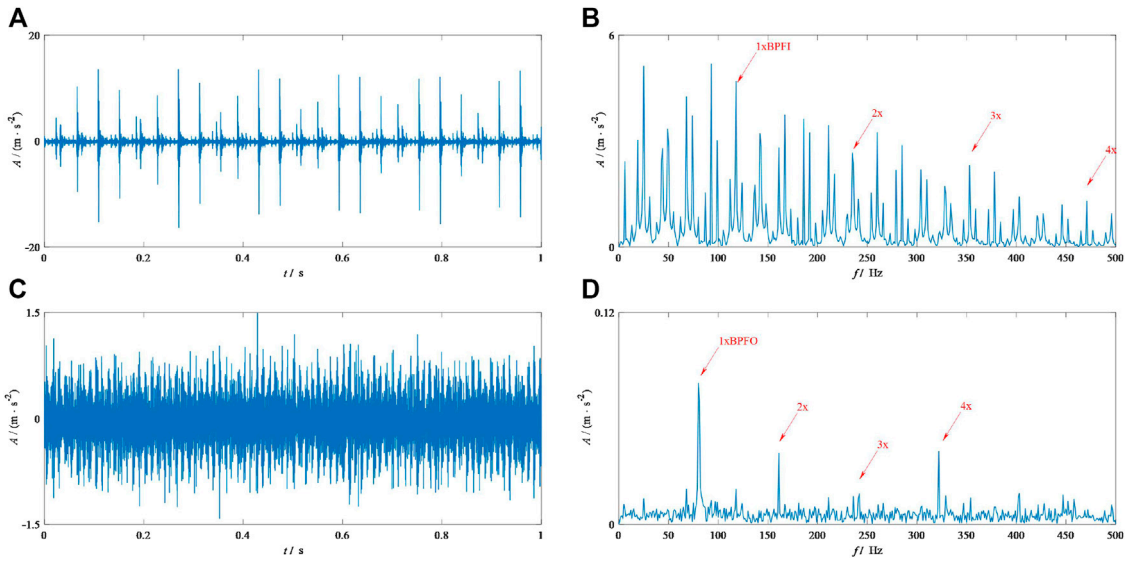


FIGURE 9
Results from ICF-VMD for the MFPT signals: **(A)** and **(B)** correspond to the TDW and the corresponding SES of the inner race fault bearing; **(C)** and **(D)** correspond to the TDW and the corresponding SES of the outer race fault bearing.

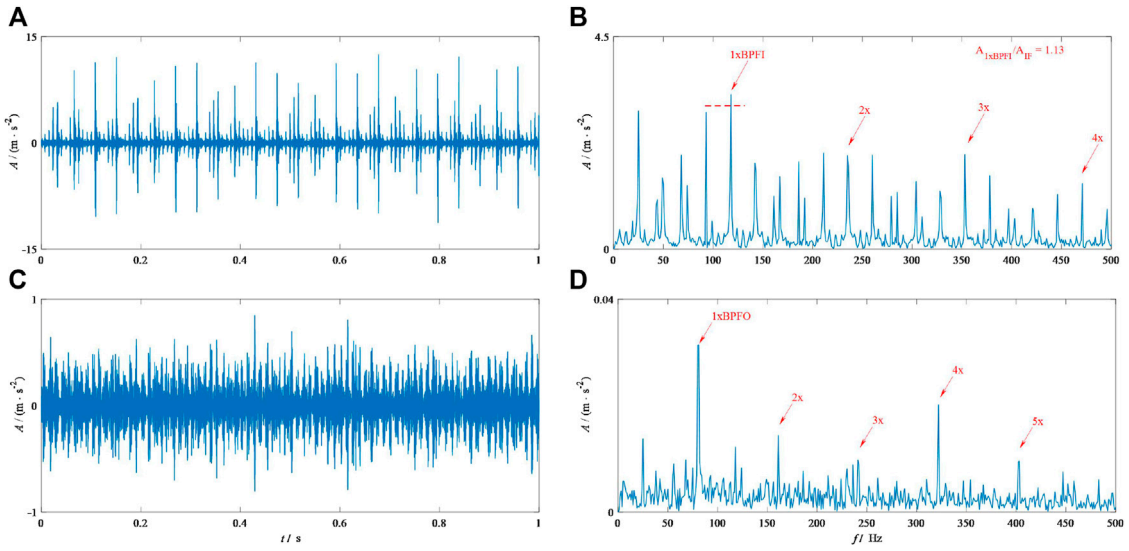


FIGURE 10
Results from IEFI–VMD for the MFPT signals: **(A)** and **(B)** correspond to the TDW and the corresponding SES of the inner race fault bearing; **(C)** and **(D)** correspond to the TDW and the corresponding SES of the outer race fault bearing.

described as extremely low. Figures 5C,D show the signal of the inner race fault bearing. Moreover, some periodic impulses can be easily found in Figure 5C. Moreover, some information about the inner race fault can be easily found in its SES as shown in Figure 5D, but some interferences occur in it. Figures 5E, F show the information about the signal of the outer race fault. Unfortunately, it is difficult to find the periodic impulses. Nonetheless, the 1xBPFO and 2x can be clearly observed in it.

Finally, this study calculates the ratio of the amplitudes between the interference frequency (IF) and fundamental feature frequency to show the superiority of the proposed method conveniently. A large value of the ratio means a good result for extracting fault features. We applied this ratio in the results of the inner race fault signal.

First of all, IEFI–MVMD is applied to process these signals. Figure 6 shows the results about the IEFI, whereas Figure 6 presents the results of the proposed method. According to Figure 5, the initial center frequencies for the healthy bearing, the inner fault bearing, and outer fault bearing should be 6,103, 3,051, and 1,907 Hz, respectively. From Figure 7B, the amplitude for either BPFO

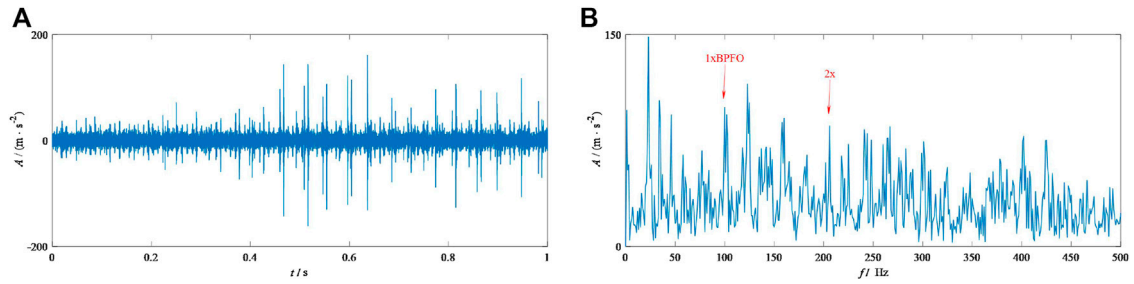


FIGURE 11
Raw signals from CU-O: **(A)** TDW and **(B)** SES.

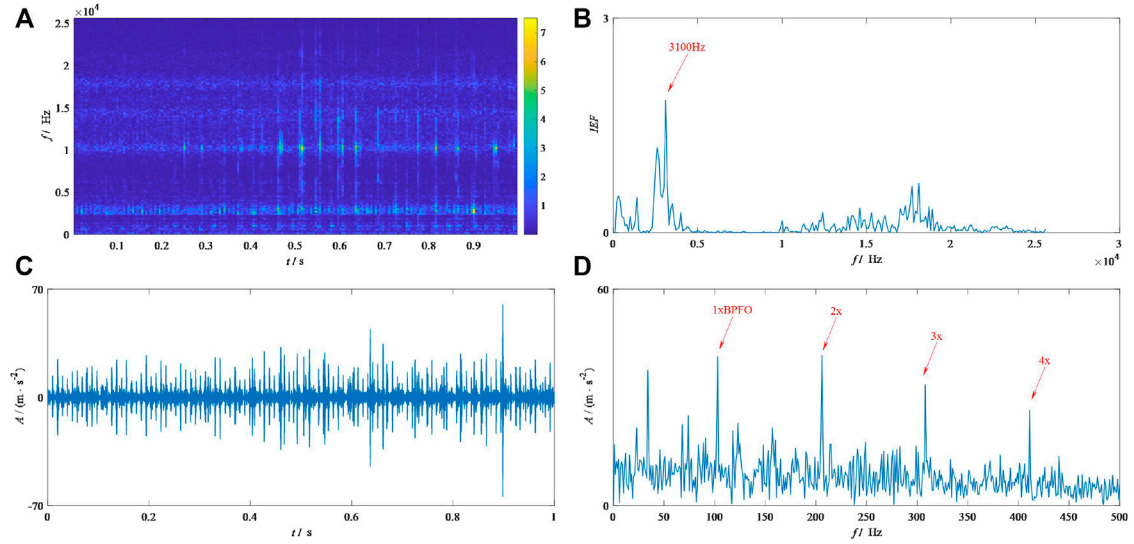


FIGURE 12
Results from IEFI–MVMD: **(A)** STFT, **B)** IEFI, **(C)** TDW, and **(D)** SES.

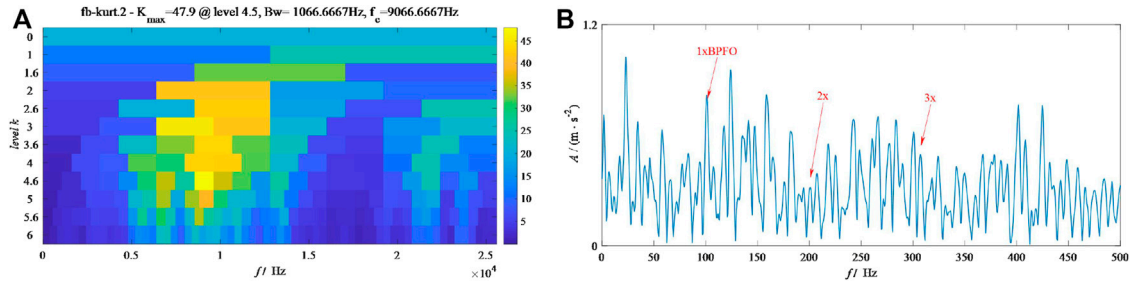


FIGURE 13
Results from FK for signal CU-O: **(A)** kurtogram and **(B)** SES.

or BPFI is low, which means the bearing is healthy. The results of the healthy bearing indicate that our method can accurately deal with these kinds of signals. Figure 7D shows the SES of the results by IEFI–MVMD for the inner race fault bearing signal. Carefully comparing it with Figure 4D, the ratio shown in Figure 7D is 1.14, which is larger than the ratio shown in Figure 4D. This finding means that IEFI–MVMD enhances the inner race fault feature. Figure 7F illustrates the SES of the results by IEFI–MVMD for the outer race fault bearing signal. Comparing it to Figure 4C, the fault feature is enhanced by IEFI–MVMD efficiency because the high-order harmonics (3x, 4x, and 5x) can only be found in Figure 6F. Based on the aforementioned description, our method can be said to reflect

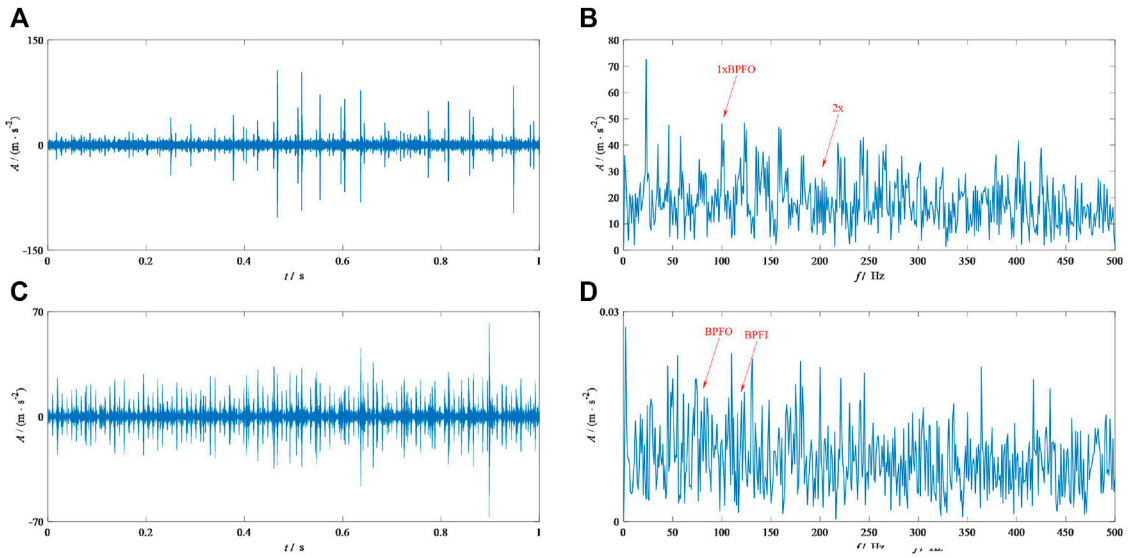


FIGURE 14
Results from signal CU-O: **(A)**ICF-VMD TDW, **(B)** ICF-VMD SES, **(C)** IEFI–VMD TDW, and **(D)** IEFI–VMD SES.

TABLE 1 Calculation time for each signal unit: (s).

Signal	IEFI–MVMD	IEFI–VMD	ICF-VMD
MFPT-H	2.17	9.97	36.69
MFPT-O	7.07	40.86	53.57
MFPT-I	1.14	7.38	16.04
CU-O	2.77	6.30	19.88

the real health status, including the health, inner race fault, and outer race fault.

The signals of inner and outer race fault bearings are analyzed by some other methods, including the fast kurtogram (FK), the ICF-VMD, and a new method that combines the IEFI and the raw VMD. For convenience, this method is named IEFI–VMD.

Figure 8 shows the results from FK. From Figure 8A, the optimal demodulation frequency band (ODFB) by FK for MFPT-I is in level 1 with the center frequency 6,103 Hz. In addition, Figure 8B shows the SES of the signal based on this ODFB. From Figure 9B, the fault features including 1 x BPFI, 2x, and 3x are clearly shown. However, the ratio shown in its upward right corner is lower than the result shown in Figure 7D, which means that our method works better than FK to extract the inner fault features. Moreover, Figure 8C shows that the ODFB by FK for MFPT-O is located in level 6 with the center frequency 17,929 Hz. In addition, Figure 8D illustrates the SES of the signal based on this ODFB. By comparing Figure 8D with Figure 5F, the high-order harmonics (4x and 5x) can only be easily found in Figure 8F. Thus, FK cannot catch up with the level of our method in dealing with both the inner and outer race fault signals.

Figure 9 and Figure 10 show the results from ICF-VMD and IEFI–VMD, respectively. Figure 9A shows the TDW of the results by ICF-VMD for MFPT-I, and Figure 9B shows its SES. By comparing Figures 8B, 6D, using the proposed method to diagnose faults provides better performance than using ICF-VMD because the amplitude of 1xBPFI is not the highest in SES and other interferences exist in it. Figure 9B displays the TDW of the results by ICF-VMD for MFPT-O, and Figure 9D shows the corresponding SES. By comparing Figures 9D, 7F, determining that the high-order fault features (3x and 5x) are weaker than the results is not difficult, as shown in Figure 7D. Figure 10A shows the TDW of the results by IEFI–VMD for MFPT-I, and Figure 10B shows its corresponding SES. By comparing Figures 10B, 7D, determining the differences between them is difficult. The ratio shown in the upward right corner of Figure 10B tells us that our method has a slight lead. Figure 10C shows TDW of the result by IEFI–VMD for MFPT-O, and Figure 10D shows its SES. From Figure 10D, some interference occurs near the fault feature 3x. Nonetheless, in Figure 7F, it is shown clearly. This finding means that the proposed method has a slight lead. More importantly, the computation efficiencies of IEFI–MVMD and IEFI–VMD are

highly different, and we will show it toward the end of this paper to highlight the superiority of our introduced method.

3.2 Case II

This section applies the introduced method to analyze another fault bearing signal, which includes some interference impulses. This kind of signal effectively highlights the advantage of our method.

The signal comes from Curtin University. The type of the test bearing is MB ER-16K, and a local defect exists in its outer race. For a convenient description, this signal is marked as CU-O in this research study. The shaft speed is 1,740 rpm, and the BPFO is 103.6 Hz from Ref. 31. The sampling frequency is 51.2 kHz, and the length of the signal applied in this study is 1 s.

Figure 11 shows the TDW and its SES. From Figure 11A, some certain interference impulses (marked by red point) exist in the measured signal. In Figure 11B, determining the fault feature frequency and its harmonics is difficult due to the interference from noise. Then, our method is applied to analyze this signal, and Figure 12 shows the results. From Figure 12A, the center frequency of the interference impulses is near 10 kHz. However, the result of IEFI shown in Figure 12B tells us that the center frequency of the periodic impulses should be 3,100 Hz, and the value of the interference impulses is extremely low. This result means that IEFI can effectively suppress the interference impulses. Figure 12C shows the TDW of the result by our method for CU-O. According to Figure 12C, some periodic impulses are clearly shown and the interference impulses are suppressed effectively. Figure 12D shows its SES. The fault features including 1xBFO, 2x, 3x, and 4x are clearly shown in it. Consequently, our method can be said to have succeeded in detecting the bearing fault feature accurately and is strong enough to resist the interference from the aperiodic impulses.

Signal CU-O is also processed by FK, ICF-VMD, and IEFI–VMD. In addition, Figure 13 and Figure 14 show their results, respectively. From Figure 13A, the ODFB FK can be seen at level 4.5 with the center frequency of 9,066 Hz. Additionally, according to SES from Figure 13B, only the fundamental fault feature frequency can be observed easily. Evidently, a large gap exists between Figures 13B, 12D. Figure 14C shows the TDW by ICD-VMD. According to Figure 14C, some interference impulses remain included in the filtered signal. Moreover, based on its SES shown in Figure 14D, observing the fundamental fault feature frequency and its harmonics is difficult due to the existence of noise and interference impulses. Figure 14C shows the TDW by IEFI–VMD for signal CU-O, and Figure 14D shows its SES. From Figure 14C, determining that the interference impulses are suppressed effectively is easy. Subsequently, according to SES from Figure 14D, the fundamental fault feature frequency and its harmonics can be observed clearly. By comparing it with the result shown in Figure 13D, we think they have the same level. However, the computational efficiency of IEFI–VMD is much farther from IEFI–MVMD.

To obtain the calculation time of IEFI–MVMD, IEFI–VMD, and ICF-VMD accurately, each method is tested three times in the same computer whose hardware is Intel(R) Core (TM) i7-9700 CPU @ 3.00 GHz 3.00 GHz. The mean is applied to evaluate the computational efficiency. Table 1 shows the results. From this table, the calculation time of our method is the lowest for each signal, which means our method has the highest computational efficiency among the three methods. Consequently, IEFI–MVMD can detect the bearing fault feature with great computational efficiency.

4 Conclusion

This study proposes a novel method named IEFI–MVMD to detect the fault feature of the bearing. IEFI–MVMD has a strong power to resist interference from aperiodic impulses and has high computational efficiency. Specifically, the guide-center frequency is determined by the IEFI calculated based on the subscript of the elements. If it is greater than the mean, the ability to resist random impulses could be enhanced. The fault feature is extracted by the MVMD whose convergence condition is built up by decomposing kurtosis, which ensures that the proposed method has high computational efficiency. The proposed method succeeds in analyzing signals from inner and outer race fault bearings and healthy bearings. The advancement of the proposed method is highlighted by comparing it to other existing methods.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: https://www.mfpt.org/fault-data-sets/.

Author contributions

Funding acquisition: YL and YM; project administration: YL and YM; conceptualization: YL and XC; validation: YL and YC; formal analysis: YL and XC; investigation: YL and XC; data curation: YL and XC; writing—original draft preparation: YL; and writing—review and editing: YL. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. U2034209), the Postdoctoral Science Foundation of China (Grant No. 2021M700590), and the Fundamental Research Funds for the Central Universities (Grant No. 2022CDJJMRH-008).

Acknowledgments

The authors would like to thank all the people who participated in the studies.

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

References

1. He B, Huang Y, Wang D, Yan B, Dong D. A parameter-adaptive stochastic resonance based on whale optimization algorithm for weak signal detection for rotating machinery. *Measurement* (2019) 136:658–67. doi:10.1016/j.measurement.2019.01.017

2. 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

3. Wang Y, Qi G, Li S, Chai Y, Li H. Body part-level domain alignment for domain-adaptive person re-identification with transformer framework. *IEEE Trans Inf Forensics Security* (2022) 17:3321–34. doi:10.1109/tifs.2022.3207893

4. Fan L, Chai Y, Chen X. Trend attention fully convolutional network for remaining useful life estimation. *Reliability Eng Syst Saf* (2022) 225:108590. doi:10.1016/j.ress.2022.108590

5. Liu B, Chai Y, Liu Y, Huang C, Wang Y, Tang Q. Industrial process fault detection based on deep highly-sensitive feature capture. *J Process Control* (2021) 102:54–65. doi:10.1016/j.jprocont.2021.04.003

6. Liu B, Chai Y, Huang C, Fang X, Tang Q, Wang Y. Industrial process monitoring based on optimal active relative entropy components. *Measurement* (2022) 197:111160. doi:10.1016/j.measurement.2022.111160

7. Liu B, Chai Y, Jiang Y, Wang Y. Industrial Fault detection based on discriminant enhanced stacking auto-encoder model. *Electronics* (2022) 11(23):3993. doi:10.3390/electronics11233993

8. 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* (2022) 206:112346. doi:10.1016/j.measurement.2022.112346

9. Yu J, Hu T, Liu H. A new morphological filter for fault feature extraction of vibration signals. *IEEE Access* (2019) 7:53743–53. doi:10.1109/access.2019.2912898

10. Zhang H, He Q. Tacholess bearing fault detection based on adaptive impulse extraction in the time domain under fluctuant speed. *Meas Sci Tech* (2020) 31(7):074004. doi:10.1088/1361-6501/ab7dec

11. Yan R, Gao RX, Chen X. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Processing* (2014) 96:1–15. doi:10.1016/j.sigpro.2013.04.015

12. Dragomiretskiy K, Zosso D. Variational mode decomposition. *IEEE Transactions Signal Processing* (2013) 62(3):531–44. doi:10.1109/tsp.2013.2288675

13. Wang Y, Markert R. Filter bank property of variational mode decomposition and its applications. *Signal Process.* (2016) 120:509–21. doi:10.1016/j.sigpro.2015.09.041

14. Wang Y, Markert R, Xiang J, Zheng W. Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system. *Mech Syst Signal Process* (2015) 60:243–51. doi:10.1016/j.ymssp.2015.02.020

15. Li Z, Jiang Y, Guo Q, Hu C, Peng Z. Multi-dimensional variational mode decomposition for bearing-crack detection in wind turbines with large driving-speed variations. *Renew Energ* (2018) 116:55–73. doi:10.1016/j.renene.2016.12.013

16. Li F, Li R, Tian L, Chen L, Liu J. Data-driven time-frequency analysis method based on variational mode decomposition and its application to gear fault diagnosis in variable working conditions. *Mech Syst Signal Process* (2019) 116:462–79. doi:10.1016/j.ymssp.2018.06.055

17. Li Y, Li G, Wei Y, Liu B, Liang X. Health condition identification of planetary gearboxes based on variational mode decomposition and generalized composite multi-scale symbolic dynamic entropy. *ISA Trans* (2018) 81:329–41. doi:10.1016/j.isatra.2018.06.001

18. Huang Y, Lin J, Liu Z, Wu W. A modified scale-space guiding variational mode decomposition for high-speed railway bearing fault diagnosis. *J Sound Vibration* (2019) 444:216–34. doi:10.1016/j.jsv.2018.12.033

19. Xu B, Zhou F, Li H, Yan B, Liu Y. Early fault feature extraction of bearings based on Teager energy operator and optimal VMD. *ISA Trans* (2019) 86:249–65. doi:10.1016/j.isatra.2018.11.010

20. Zhang X, Miao Q, Zhang H, Wang L. A parameter-adaptive VMD method based on grasshopper optimization algorithm to analyze vibration signals from rotating machinery. *Mech Syst Signal Process* (2018) 108:58–72. doi:10.1016/j.ymssp.2017.11.029

21. Miao Y, Zhao M, Lin J. Identification of mechanical compound-fault based on the improved parameter-adaptive variational mode decomposition. *ISA Trans* (2019) 84:82–95. doi:10.1016/j.isatra.2018.10.008

22. Zhao X, Wu P, Yin X. A quadratic penalty item optimal variational mode decomposition method based on single-objective salp swarm algorithm. *Mech Syst Signal Process* (2020) 138:106567. doi:10.1016/j.ymssp.2019.106567

23. Diao X, Jiang J, Shen G, Chi Z, Wang Z, Ni L, et al. An improved variational mode decomposition method based on particle swarm optimization for leak detection of liquid pipelines. *Mech Syst Signal Process* (2020) 143:106787. doi:10.1016/j.ymssp.2020.106787

24. Li Z, Chen J, Zi Y, Pan J. Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive. *Mech Syst signal Process* (2017) 85:512–29. doi:10.1016/j.ymssp.2016.08.042

25. Lian J, Liu Z, Wang H, Dong X. Adaptive variational mode decomposition method for signal processing based on mode characteristic. *Mech Syst Signal Process* (2018) 107:53–77. doi:10.1016/j.ymssp.2018.01.019

26. Wang J, Zhan C, Li S, Zhao Q, Liu J, Xie Z. Adaptive variational mode decomposition based on Archimedes optimization algorithm and its application to bearing fault diagnosis. *Measurement* (2022) 191:110798. doi:10.1016/j.measurement.2022.110798

27. Liu Y, Yang G, Li M, Yin H. Variational mode decomposition denoising combined the detrended fluctuation analysis. *Signal Process.* (2016) 125:349–64. doi:10.1016/j.sigpro.2016.02.011

28. Jiang X, Wang J, Shi J, Shen C, Huang W, Zhu Z. A coarse-to-fine decomposing strategy of VMD for extraction of weak repetitive transients in fault diagnosis of rotating machines. *Mech Syst Signal Process* (2019) 116:668–92. doi:10.1016/j.ymssp.2018.07.014

29. Gong T, Yuan X, Yuan Y, Lei X, Wang X. Application of tentative variational mode decomposition in fault feature detection of rolling element bearing. *Measurement* (2019) 135:481–92. doi:10.1016/j.measurement.2018.11.083

30. Jiang X, Shen C, Shi J, Zhu Z. Initial center frequency-guided VMD for fault diagnosis of rotating machines. *J Sound Vibration* (2018) 435:36–55. doi:10.1016/j.jsv.2018.07.039

31. Qin Y, Jin L, Zhang A, He B. Rolling bearing fault diagnosis with adaptive harmonic kurtosis and improved bat algorithm. *Ieee Trans Instrumentation Meas* (2020) 70:1–12. doi:10.1109/tim.2020.3046913

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 09 February 2023
DOI 10.3389/fphy.2023.1134431

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Jing Bi,
Guizhou University, China
Jinpeng Zhang,
Shandong Agricultural University, China
Jian Sun,
Southwest University, China

*CORRESPONDENCE
Xinrong Liu,
liuxrong@126.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 30 December 2022
ACCEPTED 27 January 2023
PUBLISHED 09 February 2023

CITATION
Guo H, Yan Y, Ding H, Liu X and Yang M (2023), Development and application of automatic monitoring equipment for differential deformation of element joint in immersed tunnel.
Front. Phys. 11:1134431.
doi: 10.3389/fphy.2023.1134431

Development and application of automatic monitoring equipment for differential deformation of element joint in immersed tunnel

Hongyan Guo[1,2], Yu Yan[3], Hao Ding[2], Xinrong Liu[1]* and Men Yang[2]

[1]School of Civil Engineering, Chongqing University, Chongqing, China, [2]China Merchants Chongqing Communications Technology Research and Design Institute Co, Ltd., Chongqing, China, [3]HZMB Administrative Authority, Zhuhai, Guangdong, China

The reliability of the immersed tunnel element joint is the key to determine whether the immersed tunnel can operate safely. At present, the immersed tunnel monitoring mostly pays attention to the joint opening and closing amount and neglects the differential deformation of the joint. Based on the immersed tunnel of Hong Kong-Zhuhai-Macao Bridge, combined with the operating environment and structural characteristics of the immersed tunnel, this paper introduces a close-range photogrammetry method to monitor the differential deformation of the immersed tunnel element joint. Through theoretical analysis, software and hardware development, laboratory test and field test, the paper puts forward puts forward a comprehensive multi-parameter evaluation and screening algorithm of boundary fitting ellipse based on fitting rate, ellipticity and area difference and a micro-displacement correction algorithm for camera based on three-dimensional calibration object, and develops an automatic monitoring system equipment for differential deformation of immersed tunnel element joint. Upon tests in tunnels, the monitoring equipment is proven in automatic monitoring on differential deformation of immersed tunnel element joints. This equipment has been successfully applied to the E31~E32 element joint of Hong Kong-Zhuhai-Macao Bridge immersed tunnel, which verifies the effectiveness of the equipment from the perspective of practical engineering application.

KEYWORDS

immersed tunnel, photogrammetry, element joint, differential deformation, automatic monitoring

1 Introduction

With the continuous development in social and economic level and the continuous improvement of people's requirements for quality of life, the overall scale of China's transportation engineering construction shows a growing trend. Tunnel, serving as engineering works of underpass, has demonstrated some incomparable advantages over other projects. As a form of underwater tunnel, immersed tunnel is increasingly favored by the engineering circle and becomes the choice of more and more cross-river and cross-sea channels. The successful construction of Subsea Immersed Tunnel of Hong Kong-Zhuhai-Macao Bridge is a herald of the maturing of the construction technology of

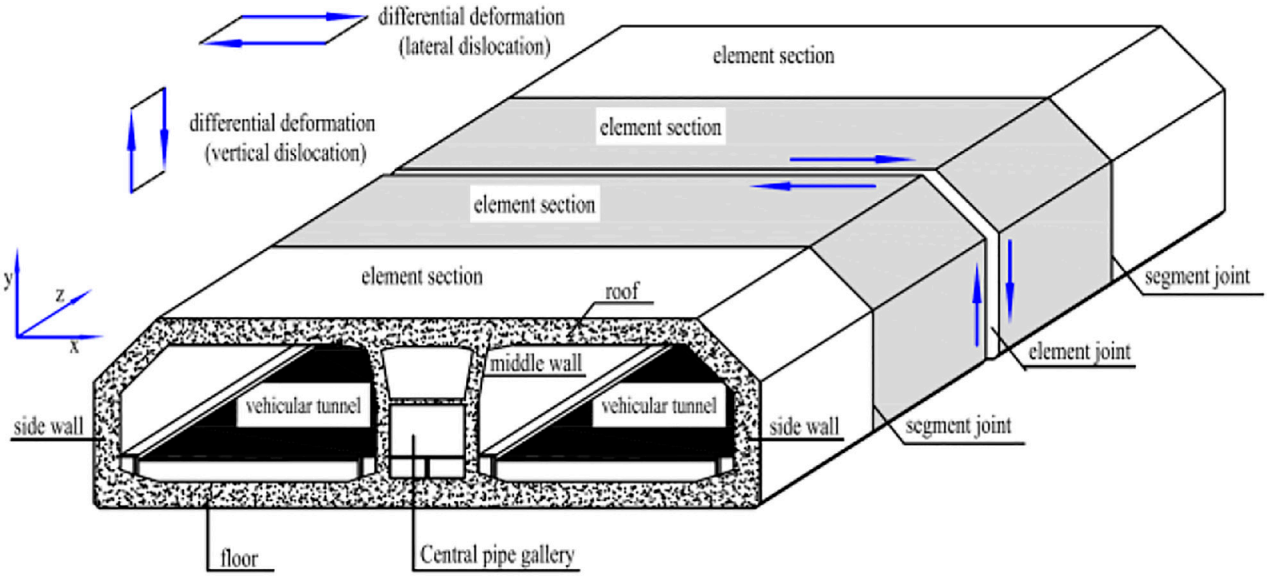


FIGURE 1
Schematic diagram of the differential deformation mode between the immersed tunnel model and the element joints.

the immersed tunnels. However, massive construction will undoubtedly bring challenges to long-term operation safety supervision, which makes the high-accuracy and non-contact automated monitoring technology research and equipment development increasingly attract people's attention.

Research results at home and abroad show that the reliability of immersed tunnel element joints is the key factor that determines the safe operation of immersed tunnels. According to the analysis on the investigation of immersed tunnels in service and literature data, the main diseases of immersed tunnels can be divided into three categories: main structural diseases, joint diseases and auxiliary structural diseases. In terms of the causes of disease, the differential deformation of element joints (see Figure 1) is the main cause of the element structural cracks and joint water damage [1], so element joints will be the focus of the operation safety monitoring of immersed tunnels.

Relatively mature monitoring methods currently for structural deformation of operating tunnels mainly include manual level monitoring, automatic static level monitoring, automatic vibrating string or optical fiber displacement meter monitoring, and automatic total station monitoring. Thanks to the progress of monitoring technology, new technologies represented by 3D laser scanning, ultrasonic sensing technology and distributed optical fiber have gradually emerged, which is a great impetus to the technical progress of tunnel structural deformation monitoring. For example: Yang Hao and Xu Xiangyang [2], et al. proposed a deformation monitoring algorithm based on laser technology, which can effectively improve the reliability of structural health monitoring; Du Liming [3] et al. developed a 3D laser scanning-based mobile tunnel monitoring system on the basis of in-depth study of tunnel monitoring methods, with a cross-section measurement accuracy of 1.1 mm; Xu Dongsheng et al. [4] developed an automatic and wireless tunnel deformation monitoring system based on ultrasonic sensing technology; Hou Gongyu [5] et al. proposed a tunnel section deformation sensing method based on distributed optical fiber sensing and neural network. In term of deformation monitoring of immersed tunnels, the immersed tunnel in Yongjiang, Ningbo, Zhejiang Province measure the overall settlement of the tunnel with levels, and then calculates the differential settlement between elements through the settlement between adjacent measuring points [6, 7]. Yuan Zheng [8] and others introduced the trilateration network method into the joint deformation monitoring of Ningbo Changhong Immersed Tunnel, realizing the monitoring of joint relative displacement. The automatic total station method was adopted by Hong Kong's first cross-sea immersed highway tunnel [9], realizing the automatic monitoring of tunnel structural displacement. In the immersed tunnel of Hong Kong-Zhuhai-Macao Bridge, a structural health monitoring system was arranged at the beginning of the construction, which realizes the automatic monitoring of the opening and closing amount of 32 element joints through fiber grating displacement meters, while the uneven settlement of the joints is still made by manual monitoring with level. In summary, the current displacement monitoring methods for tunnel structures can be roughly divided into two categories, one is the point contact measurement represented by static level and displacement meter, and the other is the optical non-contact measurement represented by automatic total station (measuring robot). The former has high measurement accuracy, but one sensor can only achieve the measurement of one displacement index of one measuring point, and the monitoring accuracy is greatly affected by the equipment installation quality. The latter can achieve simultaneous measurement of multiple measuring points and multiple displacement indexes, but the measuring point accuracy is generally at millimeter level, which cannot

meet the requirements for the observation of the small deformation of tunnel structure. In addition, measuring robots have problems such as high installation costs and easy off-target in long-term operation.

With the development of science and technology and the breakthrough of optical measurement technology, measurement and detection technology based on machine vision [10–12] has been widely valued by the industry, thus providing a new idea for non-contact, high-accuracy automatic monitoring of engineering structure displacement [13]. In civil engineering, relevant technical research and equipment development have emerged in the fields of structural experiment [14, 15], building structure [16, 17], bridge structure [18–20], hydraulic structure [21, 22], foundation pit [23, 24], slope engineering [25, 26] monitoring, etc. The research field mainly covers system design, equipment selection, image processing algorithm, engineering environment impact, and equipment R&D and application [27, 28]. In the field of tunnel and underground engineering, Cheng Zheng et al. [29] also proposed a new method in 2012 for long-span underground space structure monitoring based on digital image processing measurement technology, but neither corresponding monitoring equipment has been developed, nor engineering application has been formed, and few relevant studies have been reported since. So far, no mature visual measurement equipment suitable for the deformation monitoring of tunnels and underground engineering structures has been developed.

In conclusion, the displacement monitoring of tunnel and underground engineering structure is different from that of above-ground structures, and the existing monitoring technology and equipment are faced with many problems, such as complex installation, low monitoring accuracy and high application cost, which are difficult to meet the needs for tunnel deformation monitoring. The research and development of monitoring equipment based on machine vision and deep learning can better solve the above contradictions [30], but it needs to solve the problems such as the difficulty in extracting structural feature displacement under tunnel environmental conditions and the long-term stability of monitoring cameras. There is an urgent need for the development of new monitoring equipment to meet the monitoring needs of immersed tunnels, so as to provide data support for the safety assessment of tunnel structures. Therefore, taking the immersed tunnel of Hong Kong-Zhuhai-Macao Bridge as the background and combined with the operating environment and structural characteristics of the immersed tunnel, this paper introduces an automatic monitoring equipment for the differential deformation of immersed tunnel joints developed based on the basic principles of photogrammetry and image recognition, which realizes non-contact high-accuracy, high-frequency and all-around monitoring of multi-measuring points and multi-degree of freedom deformation immersed tunnel element joints, and successfully improves the measurement accuracy of traditional engineering structure displacement measurement based on optical principle to submillimeter level, further reducing the monitoring cost. The equipment has been successfully put into service in actual projects, supplementing the basic data for the intelligent simulation analysis of the immersed tunnel of Hong Kong-Zhuhai-Macao Bridge.

FIGURE 2
Imaging effect of circular reflective mark.

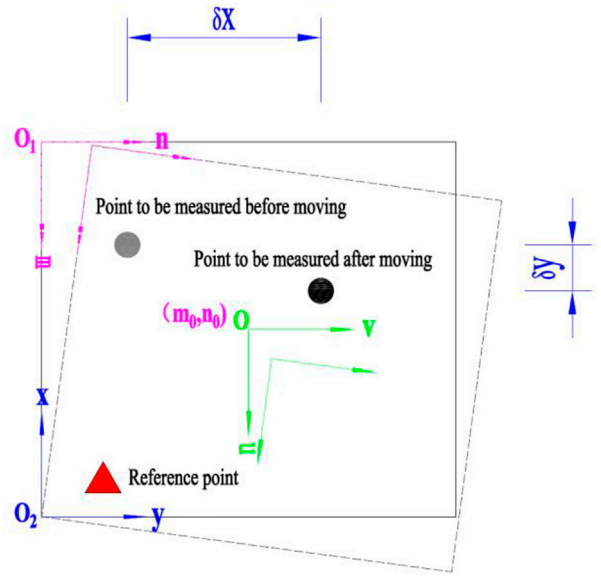


FIGURE 3
Comparison of imaging effects before and after slight changes in camera pose.

2 Measurement principle and system design

2.1 Measurement principle

For the purposes of obtaining the tunnel structure displacement, it is planned to use two “back to back” industrial cameras to obtain the 2D images of the reference point and the point to be measured respectively, see Figure 2. When the monitoring camera and the reference point are located in the stable area of the tunnel, according to the image-forming principle of cameras, if it is assumed that each image can be regarded as an array with the size of m × n, then each element of m

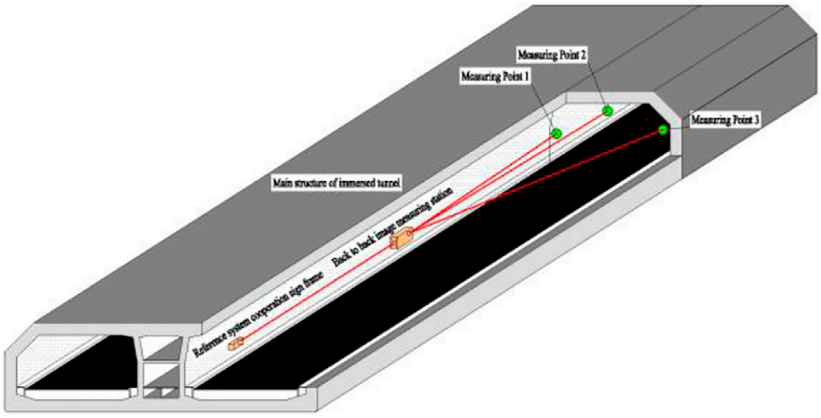


FIGURE 4
Layout and installation diagram of monitoring system.

rows and n columns corresponds to each pixel, and an O_1-mn pixel coordinate system with the upper left corner of the image as the origin of coordinates can be defined on the image, see Figure 3. To represent the pixel points in the pixel coordinate system by the method of physical quantities, it is necessary to establish an image coordinate system O-uv with the center point of the image as the origin of coordinates, see Figure 3. Through the conversion relationship between the pixel coordinate system and the image coordinate system, see Formula 1, the pixel coordinates of the point to be measured and the reference point in the image can be converted into image coordinates. In view of the fact that the images of the front and rear cameras are not in the same field of view, a reference coordinate system O_2-xy (see Figure 3) needs to be established to convert the images of the reference point and the image of the monitoring point to be monitored to a unified reference coordinate system, and then calculate the displacement change of the point to be monitored relative to the reference point (δx, δy). However, in an actual engineering environment, the change of camera installation attitudes caused by factors such as thermal expansion and contraction and structural fatigue cannot be guaranteed, but as long as the change of camera attitudes is very small, and the target to be monitored and the reference target are still located in the field of view of the cameras, the image change caused by the change of the camera attitudes will only lead to overall image displacement, and will not affect the displacement change of the point to be measured relative to the reference point, as shown in Figure 3.

$$\begin{cases} m = m_0 + \dfrac{u}{d_u} \\ n = n_0 + \dfrac{v}{d_v} \end{cases} \Rightarrow \begin{bmatrix} m \\ n \\ 1 \end{bmatrix} = \begin{bmatrix} 1/d_u & s_1 & m_0 \\ 0 & 1/d_v & n_0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (1)$$

Where, (m, n) represents the column coordinate and row coordinate of each pixel in the image; s_1 represents the non-vertical factor of the x and y axes of the imaging plane, generally 0; (m_0, n_0) represents the projection coordinates of the camera optical center on the image; d_u and d_v are the pixel sizes of the camera.

2.2 System design

The whole monitoring system is designed to be composed of three parts: the artificial cooperation reflective sign fixed in the displacement area of the tunnel to be monitored, the back to back image measuring station and the reference manual cooperative standard frame fixed in the stable area of the tunnel, as shown in Figure 4. The reference manual cooperative standard frame also provides reference coordinate system and camera attitude correction. The core equipment of the system is the back to back image measuring station, which is composed of multiple high-resolution digital cameras and camera synchronization controllers connected together. One of the cameras is equipped with a short-focus lens for imaging of the reference artificial cooperation sign frame located in the stable area of the tunnel, while the rest of the cameras are equipped with a long-focus lens for imaging of the reflective cooperation sign arranged in the deformation monitoring area of the tunnel structure in the distance. The layout of the components of the image measuring station is shown in Figure 5.

3 Structural displacement real-time monitoring technology

3.1 Sub-pixel threshold segmentation and contour extraction of monitoring point images

Since the image of the reflective sticker obtained is large and there is external environmental interference [31], in order to accurately extract the coordinates of the center of the reflective sticker, it is necessary to make *ROI* selection for the image, see Figure 6, and then conduct sub-pixel threshold segmentation and edge detection for each ROI region. In sub-pixel threshold segmentation, the first step is to perform bilinear interpolation algorithm processing for the gray value edge between two adjacent pixels on the image of the ROI region in the horizontal dimension and vertical dimensions (see Figure 7) to transformed the image of the ROI region at pixel-level accuracy to sub-pixel level

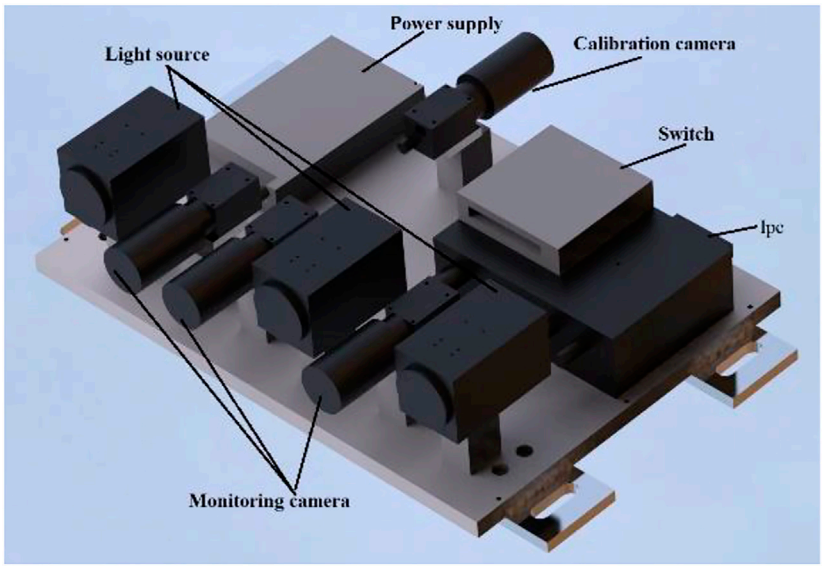


FIGURE 5
Design scheme of back to back measuring station.

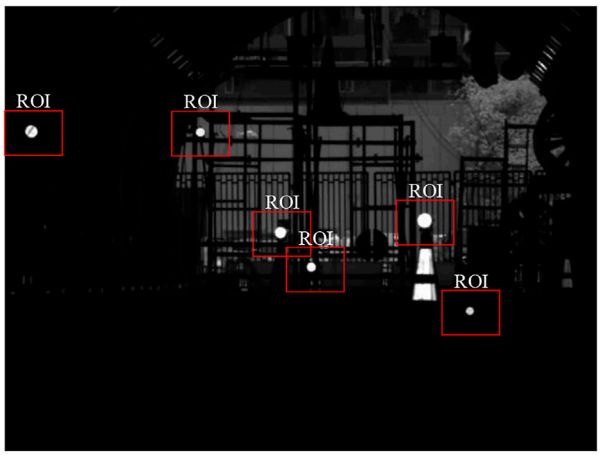


FIGURE 6
ROI selection of reflective sticker images.

accuracy [32]. After bilinear interpolation, the gray value at $p_{i+x,j+y}$ can be calculated according to Formula 2. The last step is to perform threshold segmentation according to Formula 3 to obtain the sub-pixel precision edges of the image of the ROI region for center coordinate extraction.

$$g_{i+x,j+y} \approx [1-x, x]\begin{bmatrix} g_{i,j} & g_{i,j+1} \\ g_{i+1,j} & g_{i+1,j+1} \end{bmatrix}\begin{bmatrix} 1-y \\ y \end{bmatrix} \quad (2)$$

Where, i, j represent the coordinate position of the image pixel (the value is a non-negative integer); x. y is the weight carried by the pixel value at the vertex during interpolation (this is a floating point number in the range of [0,1]); $g_{i,j}$, $g_{i+1,j}$, $g_{i,j+1}$ and $g_{i+1,j+1}$ are the gray values of four adjacent pixels an arbitrary 2 × 2 area on the image.

$$S = \left\{(r,c) \in R \,\middle|\, g_{\min} \le f_{r,c} \le g_{\max}\right\} \quad (3)$$

Where, $f_{r,c}$ is the gray value at an arbitrary pixel point (r, c) in the ROI region, g_{min} is the gray value of the pixel point at the peak on the gray histogram of the ROI region, g_{max} is the gray value of the pixel point at the peak value on the gray histogram of the ROI region, and S is the set of pixels satisfying Formula 3 on the ROI region.

3.2 Extraction of center coordinates of monitoring points based on ellipse fitting

Least square fitting [33] is the most common fitting method in ellipse fitting. However, the fitting degree between the ellipse and the original boundary is not taken into account when this method is used for ellipse fitting, and the resulting ellipse is larger than the actual one. The results of this method are often not satisfactory, which makes it necessary to evaluate and screen the fitted ellipse. In this paper, a boundary fitting ellipse multi-parameter comprehensive evaluation and screening algorithm based on fitting rate, ellipticity and area difference is proposed. Assuming that the equation of the fitted ellipse (see Figure 8) is as shown in Formula 4, P (m, n) is an arbitrary boundary point used for the fitting, and Q (x, y) is the intersection point of the normal line passing through Point P and the fitted ellipse, the coordinates of Point Q can be calculated from Equation Set (5). Traverse each boundary point and calculate the vertical distance d between the boundary points and the fitted ellipse according to Formula 6, denote the points whose vertical distance d is less than a certain threshold T_d as matching points, and define the ratio η of the number of matching points (Pm) to the total number of boundary points (Pe) involved in the fitting as the elliptic fitting rate, see Formula 7. In addition, define the ellipticity ρ to evaluate the fitting degree of the fitted ellipse towards a circle, see Formula 8, and define the area difference $\Delta Area$ to evaluate the proximity of the fitted ellipse to the ideal ellipse, see Formula 9.

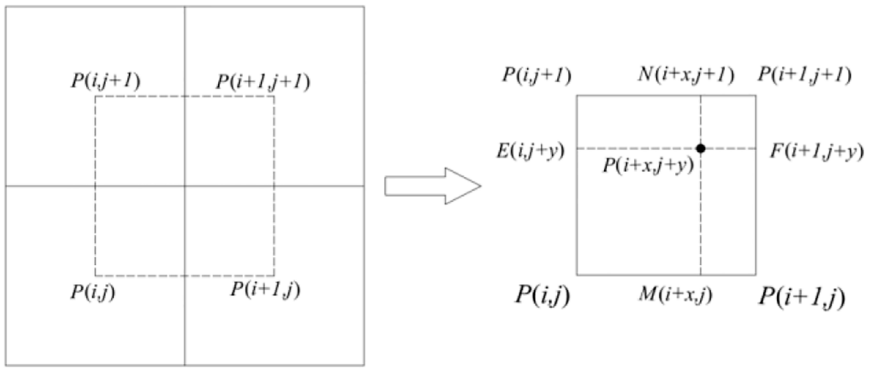


FIGURE 7
Subdivision process of pixel-level units by bilinear interpolation.

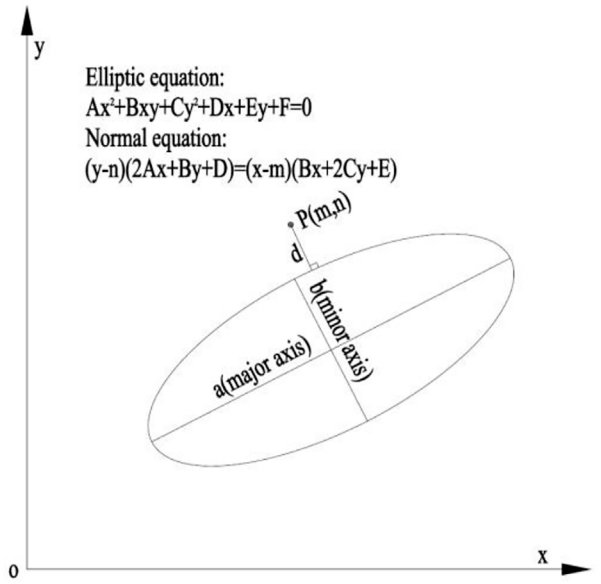


FIGURE 8
Fitting the ellipse boundary.

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \tag{4}$$

$$\begin{cases} Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \\ (n-y)(2Ax + By + D) = (m-x)(2Cy + Bx + E) \end{cases} \tag{5}$$

$$d = \sqrt{(m-x)^2 + (n-y)^2} \tag{6}$$

$$\eta = \frac{P_m}{P_e} \tag{7}$$

$$\rho = \frac{b}{a} \tag{8}$$

$$\Delta Area = \left| Area_0 - Area_f \right| \tag{9}$$

Where, A—F are the fitting coefficients of the ellipse fitting equation; a is the major axis of the fitted ellipse; b is the minor axis of the fitted

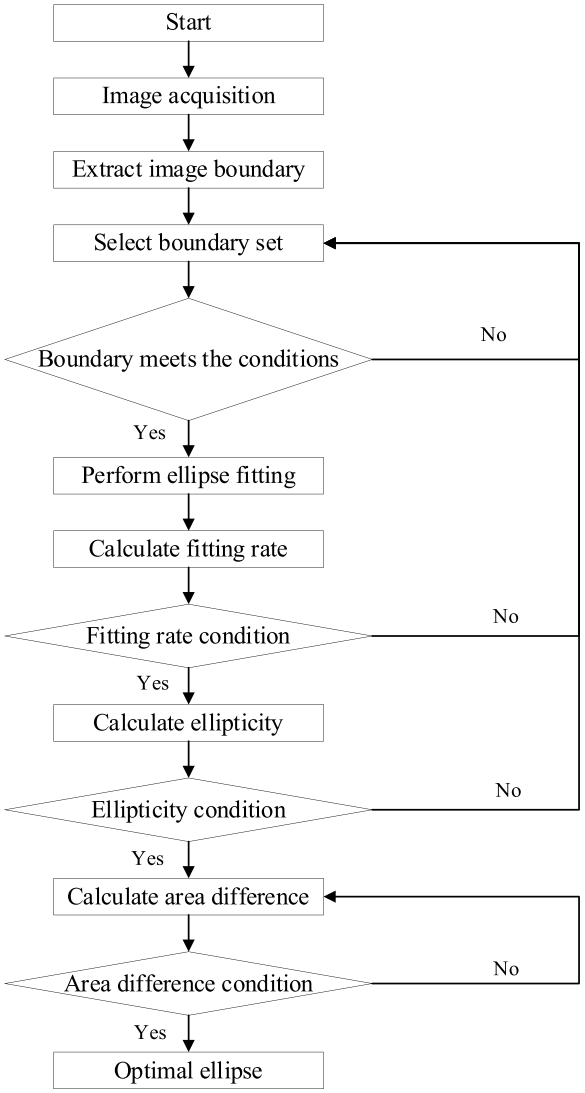


FIGURE 9
Algorithm flow chart.

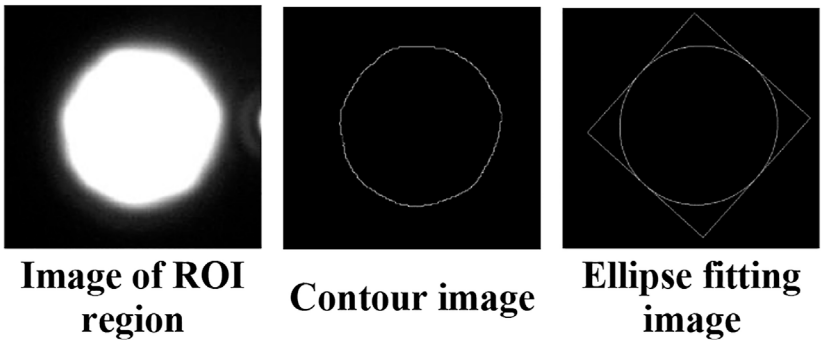


FIGURE 10
ROI area ellipse detection.

FIGURE 11
Extraction of ROI area of monitoring points.

ellipse; $Area_0$ is the area of the ellipse expected to be obtained by fitting under ideal conditions, and $Area_f$ is the area of the ellipse obtained by least square fitting under actual conditions.

The detailed realization process of boundary fitting ellipse multi-parameter comprehensive evaluation and screening algorithm based on fitting rate, ellipticity and area difference is as follows.

The boundary fitting ellipse multi-parameter comprehensive evaluation and screening algorithm based on fitting rate, ellipticity and area difference involves segmentation of image boundary, extraction of image boundary points, calculation of fitting rate and ellipticity, and selection of optimal ellipse, etc. The detailed realization process is as follows, See Figure 9.

After sub-pixel threshold segmentation and contour extraction, the center coordinates of one of the ROI regions are obtained according to the above ellipse fitting and screening algorithm. The processing effect is shown in Figure 10. As can be seen from the figure that after the ROI region is selected, we can detect the contour of the reflective sticker very well by setting the threshold, and then obtain a fitted ellipse. In the figure above, the coordinates of the center of the ellipse are (251.834991 and 219.546951), which are coordinates at the sub-pixel accuracy level.

3.3 Algorithm for transformation of monitoring point displacement in reference coordinate system

Those obtained by the above algorithm are the pixel coordinates of the points to be measured, while the structural displacement monitoring needs to obtain the displacement changes of the points to be measured relative to the reference point, so it is necessary to solve the conversion relationship between pixel coordinate system and reference coordinate system. In this case, it is necessary to complete the measurement with the aid of a total station. First, place reflective stickers at the monitoring parts of the tunnel, and use a total station to obtain the three-dimensional coordinate points P_i (i = 1, 2, 3,. . .,) of the center of each reflective sticker. Similarly, use a laser LED light to illuminate the reflective stickers. Then, collect the image of the reflective stickers, extract the ROI region in the image, obtain the reflective sticker area image (as shown in Figure 11), perform ellipse fitting for each ROI, and obtain the center coordinates (u, v). In the meantime, conduct distortion correction according to the distortion correction model shown in Formula 10, and use the corrected pixel coordinates as the initial coordinate values of the monitoring points.

$$\begin{cases} d_x = x[k_1 r + k_2 r^2 + p_1 r + p_2 r^2] \\ d_y = y[k_1 r + k_2 r^2 + p_1 r + p_2 r^2] \end{cases} \tag{10}$$

Where, d_x and d_y are distortion values, p_1 and p_2 are tangential distortion coefficients, and k_1 and k_2 are radial distortion coefficients; $r = \sqrt{x^2 + y^2}$.

Calculate the rotation R and translation t from the reference coordinate system set by the total station to the monitoring camera coordinate system through *PnP* algorithm [34] based on the corresponding relationship between the 2D coordinate point of the center of the circle and the 3D coordinate point obtained by the total station, and on this basis, transform the 3D coordinate point P_i obtained by the total station to the monitoring camera coordinate system. The position relationship between the reference coordinate

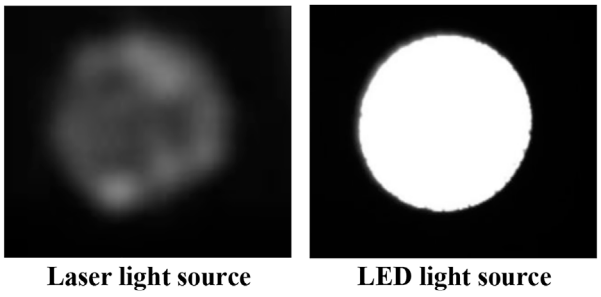


FIGURE 12
Comparison of imaging effects of different light sources.

systems of the monitoring camera and the total station is shown in Formula 11:

$$P_{ci} = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \cdot P_i \tag{11}$$

Assuming that the position attitude relationship between the two cameras obtained through external parameter calibration is T_{AB}, and the initial position of the calibration camera relative to the reference marker (A marker considered to be fixed and immovable) is H, then the three-dimensional coordinates P_{wi} of the monitoring point P_{ci} in the marker coordinate system can be calculated according to Formula 12:

$$P_{wi} = HT_{AB}P_{ci} \tag{12}$$

3.4 Correction of measurement deviation caused by small camera displacement

When there is human interference or when the tunnel where the camera is installed deviates, the initial position of the measuring point viewed by the monitoring camera will change [35]. In this case, the initial position of the monitoring point needs to be corrected with the calibrated camera parameters. Assuming that the coordinates of the monitoring point in the reference marker coordinate system are P_{wi}, and the positional relationship of the calibration camera relative to the reference marker is H_{new}, when the position information of the calibration camera changes, the position coordinates of the monitoring point P_i after correction can be calculated according to Formula 13:

$$P_i = T_{AB}H_{new}P_{wi} \tag{13}$$

4 Hardware selection and structure design

In order to meet the structural displacement measurement function in tunnel light environment, the measuring station mainly consists of industrial camera, lens, light source, IPC, switch and power supply. The industrial camera and supporting lens are mainly for image acquisition of the points to be measured, the light source can

TABLE 1 Test results of the best field of view width for different lenses.

Lens (mm)	Maximum width of field of view (m)		
	50	60	70
25	14	14	14
35	14	14	14
50	10	12	13
100	3	4	6

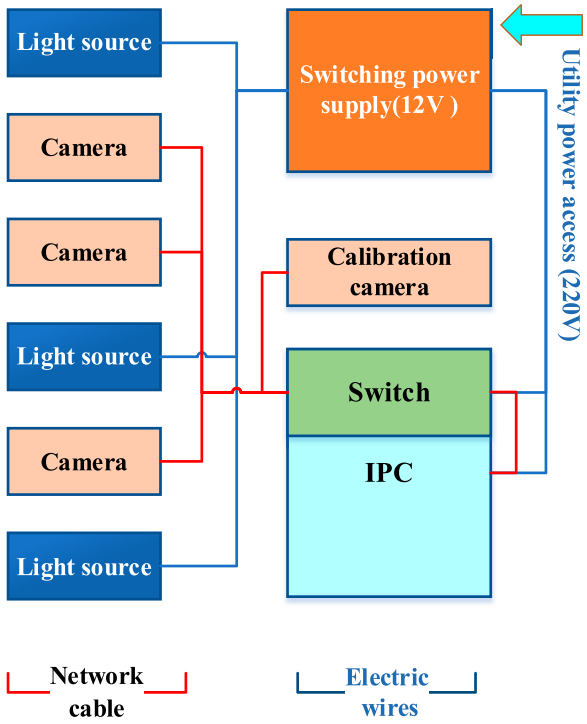


FIGURE 13
Hardware structure diagram.

provide a stable light environment for image acquisition, the switch is designed for image data transmission, and the built-in image analysis algorithm of the IPC is used to extract the monitoring results and output them externally. In terms of camera selection, since the tunnel structure displacement monitoring target is basically in a static state, the requirements for camera acquisition speed are not high, but considering that the device works outdoors, the power consumption of the device should be low. In addition, considering the popularization of the equipment, it is appropriate to select an industrial camera with CMOS chip as the image acquisition equipment. Based on the requirements of monitoring accuracy and environmental applicability, the selected industrial camera has a resolution of 2,592*2048, a photosensitive element size of 12.4 mm*9.8 mm, and a pixel size of 4.8*4.8 μm. In the aspects of light source selection, we can compare the test results of laser and LED light sources, see Figure 12. Laser fill-in light produces speckles that change with time, and such circular speckles greatly affect the accuracy

FIGURE 14
Hardware structure assembly.

of the later identification point extraction algorithm, thus affecting the detection accuracy of the algorithm. The infrared fill-in light of LED produces uniform and stable light without speckle, which is suitable for image processing algorithms. However, the power of infrared fill-in lights is weaker than that of laser lights, so multiple infrared fill-in lights are required for filling the light at the same time. When it comes to lens selection, the test results of the maximum clear imaging range of lenses with different focal lengths at different shooting distances are shown in Table 1. As can be seen from the table, the larger the focal length of the lens, the closer the shooting distance, the smaller the field of view, and when the camera pixel is fixed, the higher the accuracy of the measurement results will be. Considering accuracy, cost and actual field of view demand, 100 mm lens is selected as the final selection, when the theoretical measurement accuracy of the tested camera can reach 0.1 mm, the field of view can cover the monitoring target points in the area to be measured. The structural design and assembly of the back to back measuring station are shown in Figures 13, 14.

5 Test and analysis of equipment measurement accuracy

In order to ensure the engineering applicability of the developed automatic monitoring equipment for differential deformation of immersed tunnel joints, the accuracy and stability tests were carried out on the basis of the 200 m long 1: 1 experimental tunnel of the National Engineering Research Center for Highway Tunnels.

5.1 Test method

Three measuring points were arranged on the left, middle and right) at every 10 m interval within the range of 10–70 m from the camera, the optical mobile platforms with an accuracy of 0.01 mm were placed on the points to be measured in sequence, see Figure 15, and the camera field of view was set to 5 m. After adjusting the brightness and position of the LED, the LED light was turned on, and a filter was placed on the camera. The reflective stickers on each point to be measured were moved transversely and horizontally by 0.2, 0.3, 0.4, and 0.5 mm to simulate the small deformation of the tunnel structure. In the meantime, the camera parameters were adjusted to capture clear images, and the image processing algorithm was used to extract the center displacement of the reflective sticker images and test the measurement accuracy of the equipment at different measuring points. To ensure the accuracy of the tests, three measurements were made for each measuring point position to get an average value.

5.2 Analysis of test results

The extracted measurement error statistics under different working conditions are shown in Figure 16. As can be seen from Figure 16, when the distance between the camera and the target point was 70 m, and when the structural displacement was between 0.2 and 0.5 mm, the test error of the equipment measurement results was large, and the test data was abnormal. Therefore, the effective monitoring range of the monitoring equipment is 60 m in the longitudinal direction. A camera with a fixed focal length had a coverage of 20 m (longitudinal) * 5 m (transverse), and a minimum recognizable displacement of 0.2 mm. The measurement error gradually increased with the distance from the target point. In the same cross section, the measurement errors of different measuring points were fairly close, that is, the left or right deflection of the cameras had little impact on the measurement errors. Therefore, in practical projects, full coverage of the entire cross section can be achieved by installing monitoring equipment on the side walls of tunnels. Within the above effective measurement range, the measuring equipment had an average measurement error of 0.1 mm and a maximum measurement error of 0.19 mm. To sum up, within the effective measurement range, the equipment can identify deformations of 0.2 mm, with an average measurement error of ±0.1 mm, so it can meet the monitoring accuracy requirements of operating tunnel structures.

6 Engineering application verification

6.1 Selection of monitoring parts and installation of instruments and equipment

Differential settlement of elements and disharmonic deformation between elements in immersed tunnels are the main reasons for various diseases such as concrete structure cracks and joint leakage. Currently, the immersed tunnel of Hong Kong-Zhuhai-Macao Bridge has relatively comprehensive monitoring contents and indicators, but the monitoring frequency, monitoring methods and measuring point layout of some key

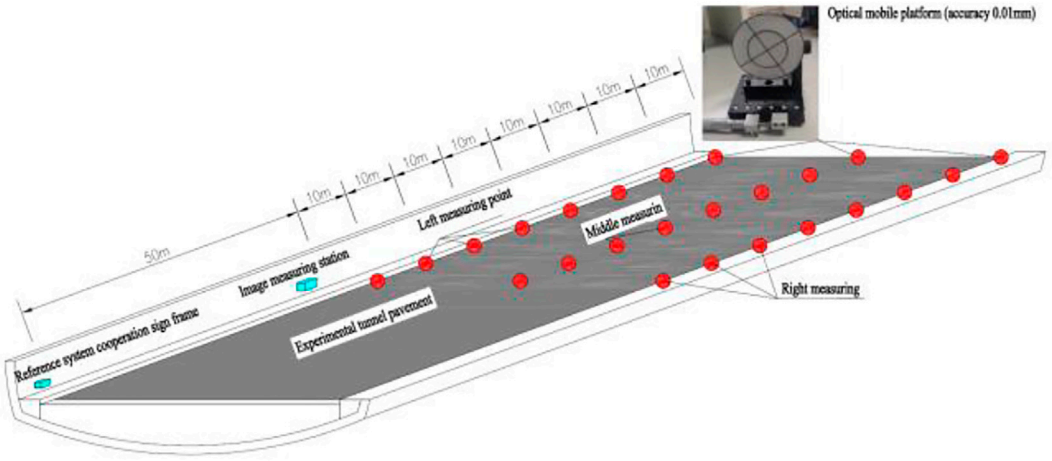


FIGURE 15
Schematic diagram of solid tunnel test.

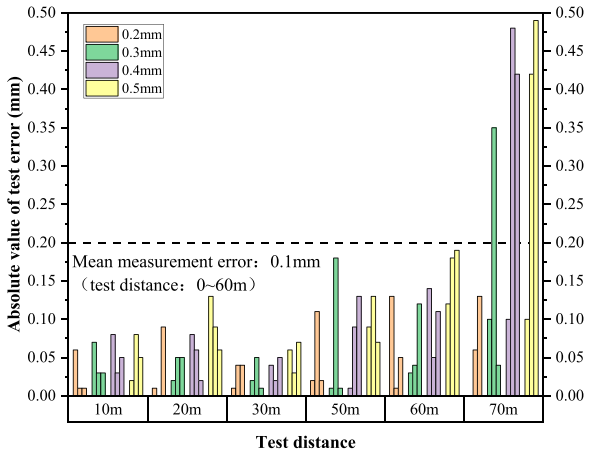


FIGURE 16
Statistical histogram of absolute value of test error.

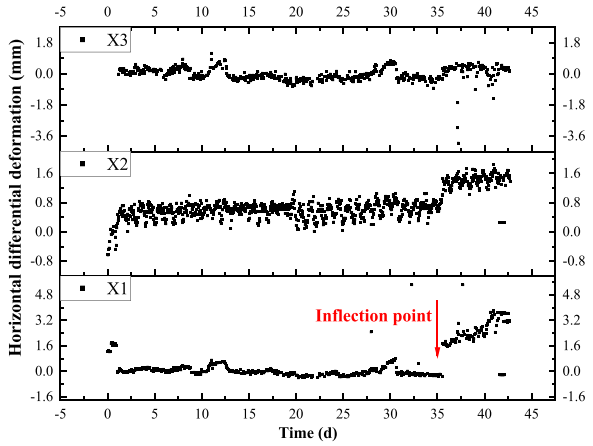


FIGURE 17
Scatter diagram of time domain analysis on monitoring results of horizontal differential deformation of joints at different measuring points.

indexes cannot meet the requirements for intelligent simulation analysis of the immersed tunnel. According to the current joint opening and closing amount and foundation settlement monitoring data, E31~E32 element joints have the largest structural displacement, and the safety risk is high. In order to meet the needs for joint safety assessment, E31~E32 element joints (Zhuhai-Macao direction) were selected to install element joint differential deformation monitoring system equipment to supplement the monitoring of the lateral and vertical differential deformation of joints. The set sampling frequency was 1 min/time. A set of reflective signs were arranged on the middle wall (Measuring Point 1), roof (Measuring Point 2) and side wall (Measuring Point 3) of the joints. The back to back measuring station was installed on the element structure above the side wall decorative plate 30 m away from the joint. The reference point was located on a stable structure 50 m away from the measuring station, and the installation height was basically consistent with the measuring station. The layout of measuring points and equipment installation are shown in Figure 4.

6.2 Monitoring results and application effect analysis

After 7 months (210 days) of testing, the long-term stability and reliability of the monitoring equipment were verified. Analysis was made on the differential deformation monitoring results of the monitored joints, and time-domain analysis (see Figures 17, 18) and boxline diagram analysis (see Figures 19, 20) were performed on the raw data extracted from the field monitoring of the last 45 days (16 April 2022–31 May 2022).

Horizontal differential deformation: As can be seen from Figure 17, the horizontal differential deformation of E31~E32 element joints can be divided into two stages. The first stage is before 20 May 2022, when it was at a relatively stable level, and the horizontal differential deformation at the measuring point of the roof showed a slight positive increasing trend. Through the combination

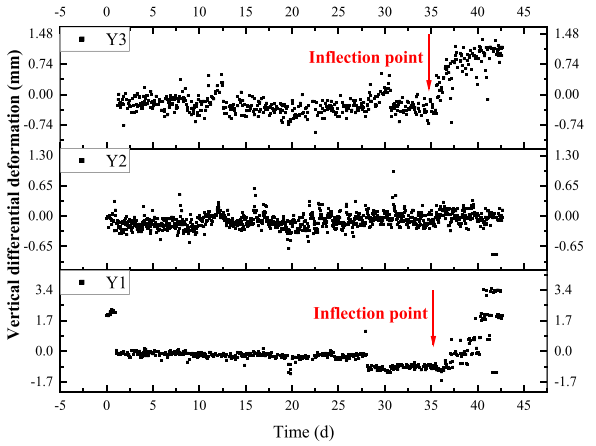


FIGURE 18
Scatter diagram of time domain analysis of monitoring results of vertical differential deformation of joints at different measuring points.

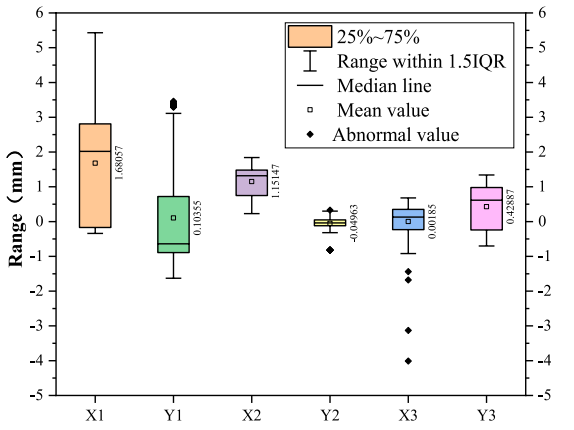


FIGURE 20
Analysis of box line analysis of joint difference deformation monitoring at different measuring points from 20 May 2022 to 31 May 2022.

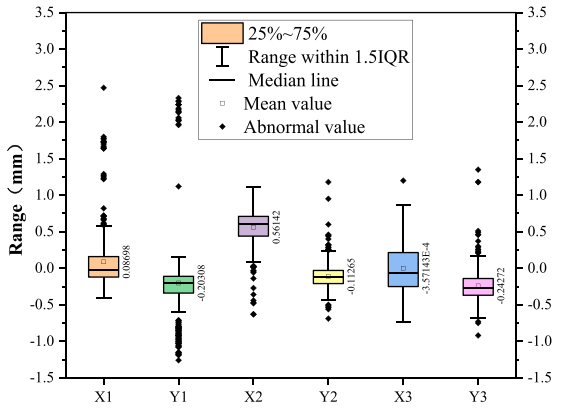


FIGURE 19
Analysis of box line analysis of joint difference deformation monitoring at different measuring points from 16 April 2022 to 20 May 2022.

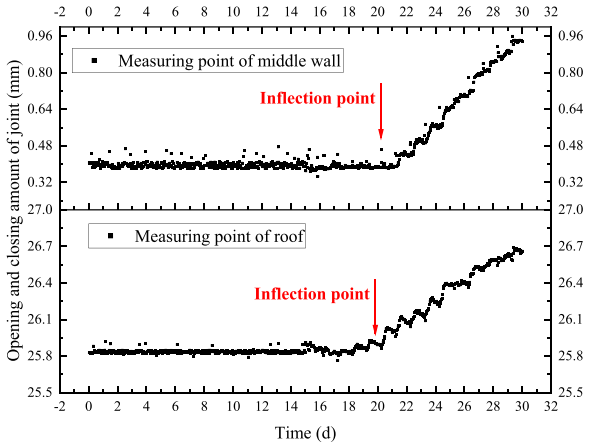


FIGURE 21
Change rule of element joint opening and closing amount at different measuring points.

with Figure 19, it can be seen that after removing the outliers, the mean values of the three measuring points basically coincided with the modes, the average value of the horizontal differential deformation (X1) at the measuring point of the middle wall was 0 mm, and the distribution of the monitoring data was mainly between −0.2 and 0.2 mm; the mean value of the horizontal differential deformation (X2) at the measuring point of the roof was 0.5 mm, and the distribution of the monitoring data was mainly between 0.4 and 0.6 mm; and the mean value of the horizontal differential deformation (X3) at the measuring point of the side wall was 0 mm, and the distribution of the monitoring data was mainly between −0.3 and 0.3 mm. At this stage, except for the 0.5 mm horizontal differential deformation accumulated at the measuring points of the roof, the horizontal deformation of the middle wall and the side wall was relatively small and stable. The fluctuation of the monitoring data was caused by the measurement error of instruments and equipment or periodic vibration noise. The second stage is from May 20 to 31 May 2022. During this period, except for the measuring point of the side wall still maintaining the original stable deformation level, the horizontal difference deformation at the measuring points of both the middle wall and the roof suddenly accelerated. The horizontal differential deformation at the measuring point of the roof experienced a sudden change, with an increase of about 1mm, and then entered the stable deformation stage again, with the deformation trend basically consistent with the previous stage, showing a slight increase trend. However, the horizontal differential deformation at the measuring point of the middle wall showed a continuous increase trend, with a deformation rate of about 0.3 mm/day. According to the comparative analysis of Figures 19, 20, the horizontal differential deformation at the measuring points of the middle wall and the roof in the two stages had a significant shift in the distribution position of boxline diagram, showing a significant positive increase, with the measuring point of the middle wall increasing by 3 mm and the measuring point of the roof increasing by 1 mm.

Vertical differential deformation: As can be seen from Figure 18, the vertical differential deformation of E31~E32 element joints can also be

divided into two stages. The first stage is before 20 May 2022, when it was at a relatively stable level, and the vertical differential deformation at the measuring point of the middle wall showed a slight negative increasing trend, and the vertical differential deformation at the measuring point of the roof had a slight positive increasing trend. Through the combination with Figure 20, it can be seen that after removing the outliers, the mean values of the three measuring points basically coincided with the modes, the average value of the vertical differential deformation (Y1) at the measuring point of the middle wall was −0.3 mm, and the distribution of the monitoring data was mainly between −0.5 mm and −0.1 mm; the mean value of the vertical differential deformation (Y2) at the measuring point of the roof was 0 mm, and the distribution of the monitoring data was mainly between −0.1 and 0.1 mm; and the mean value of the vertical differential deformation (Y3) at the measuring point of the side wall was −0.2 mm, and the distribution of the monitoring data was mainly between −0.3 and 0.1 mm. The vertical deformation of each measuring point in this stage was relatively small and stable. The fluctuation of the monitoring data was caused by the measurement error of instruments and equipment or periodic vibration noise. The second stage is from May 20 to 31 May 2022. During this period, except for the measuring point of the roof still maintaining the original stable deformation level, the vertical difference deformation at the measuring points of both the middle wall and the side wall suddenly accelerated. The vertical differential deformation at the measuring point of the side wall formed a parabolic growth, and the growth rate gradually decreased, with an increase of about 1 mm. However, the vertical differential deformation at the measuring point of the middle wall showed a trend of exponential increase, the growth rate gradually increased, with a deformation rate of about 0.25 mm/day. According to the comparative analysis of Figures 19, 20, the vertical differential deformation at the measuring points of the middle wall and the side wall in the two stages had a significant shift in the distribution position of boxline diagram, showing a significant positive increase, with the measuring point of the middle wall increasing by 2.5 mm and the measuring point of the roof increasing by 1 mm.

In order to further demonstrate the reliability of joint monitoring results, other monitoring indexes of element joints were extracted for comparative analysis. For the moment, two fiber grating displacement meters are installed at E31~E32 element joints (Zhuhai-Macao direction) to monitor the opening and closing amount of the joints, which are respectively located at the measuring point of the middle wall of the central pipe gallery and the measuring point of the roof of the vehicular tunnel. See Figure 21 for the variation law of the opening and closing amount of the joints at the two measuring points on E31~E32 element joints (Zhuhai-Macao direction) with time from 1 May 2022 to 30 May 2022. It can be seen from Figure 21 that from 20 May 2022, the opening and closing amount of the element joints also shows a certain increasing trend, which further demonstrates the reliability of the above monitoring results.

Based on the analysis of the monitoring results of the two monitoring indicators of the above two monitoring equipment, since 20 May 2022, the differential deformation of the joints and the opening and closing amount of the joints of E31~E32 element joints increased suddenly, and showed a trend of continuous increase, lasting for a long time. The preliminary judgment is that there may be abnormal changes in the environmental conditions of the element joint (differential settlement of the foundation or change in the thickness distribution of the overburden), which in turn caused lateral differential deformation and longitudinal opening and closing of the element joint. Considering that the differential deformation of the three parts measured was inconsistent, it is analyzed that there may be a certain amount of torsional deformation in the element joints. However, the GINA waterstop designed at present is controlled based on the amount of compression. Excessive shear deformation may have a greater impact on the watertightness of the joint, so there is a big potential safety hazard. Therefore, it is necessary to strengthen the monitoring of the element joint. When necessary, special detection of the foundation and the overburden and special assessment of joint safety can be carried out. In case of long-term irreversible shear deformation or sudden change of deformation value, intervention should be carried out in advance to ensure operation safety.

7 Conclusion

For the purposes of achieving the automatic monitoring of differential deformation of immersed tunnel joints, photogrammetry and image recognition technology are introduced in this paper. Based on the joint characteristics of immersed tunnels, a set of automatic monitoring equipment for differential deformation of immersed tunnel joints has been developed after a lot of efforts from technical principles to image recognition algorithms, from hardware structure design to system software development, which has solved practical engineering problems and achieved good application results.

(1) A boundary fitting ellipse multi-parameter comprehensive evaluation and screening algorithm based on fitting rate, ellipticity and area difference is proposed, and the optimal fitted ellipse boundary and monitoring point center coordinates of the image of ROI region are extracted, as improves the measurement accuracy of monitoring equipment.
(2) A camera small displacement correction algorithm based on stereo calibration object is proposed, which eliminates the measurement error caused by the change in camera attitude during the long-term operation of the monitoring equipment and solves the problem in long-term stability of the monitoring equipment.
(3) According to the test results in an actual tunnel, the effective monitoring range of the monitoring equipment is 60 m in the longitudinal direction, the minimum identifiable structural displacement is 0.2 mm, and the average measurement error is ±0.1 mm, which realizes the high-accuracy non-contact automatic monitoring of the differential deformation of immersed tunnel element joints.
(4) The test results of the application in the immersed tunnel project of Hong Kong-Zhuhai-Macao Bridge show that the equipment has a long-term stability and reliability. The equipment has successfully captured the abnormal deformation of E31~E32 element joints in the supporting project, providing effective data support for the safety assessment of the engineering structure.
(5) This equipment is still in its engineering prototype stage at present, requiring a lot of efforts in engineering application verification and optimization to ensure good engineering applicability, stability and reliability. With the progress of technology and the reduction of camera cost, this equipment is expected to have a broader application prospect.

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

HG wrote the original manuscript and carried out the hardware design and development. YY, HD, and XL proposed the idea, supervised the research work, and revised the manuscript. HG and MY discussed and analyzed the results.

Funding

The present work was supported by the National Key R&D Program of China (No. 2019YFB1600702).

Conflict of interest

Authors HG, HD, and MY were employed by China Merchants Chongqing Communications Technology 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. Li H, Jiang G, Zhang G, et al. Consideration on common diseases of immersed tunnel and its inspection and maintenance. *Tunnel Construction* (2019) 39(A01):6. doi:10.3973/j.issn.2096-4498.2019.S1.063

2. Yang H, Xu X. Structure monitoring and deformation analysis of tunnel structure. *Compos Structures* (2021) 276:114565. doi:10.1016/j.compstruct.2021.114565

3. Du L, Zhong R, Sun H, Zhu Q, Zhang Z. Study of the integration of the CNU-TS-1 mobile tunnel monitoring system. *Sensors* (2018) 18(2):420. doi:10.3390/s18020420

4. Xu DS, Zhao YM, Liu HB, Zhu HH. Deformation monitoring of metro tunnel with a new ultrasonic-based system. *Sensors* (2017) 17(8):1758. doi:10.3390/s17081758

5. Hou G, Li Z, Hu Z, Feng D, Zhou H, Cheng C. Method for tunnel cross-section deformation monitoring based on distributed fiber optic sensing and neural network. *Opt Fiber Techn* (2021) 67:102704. doi:10.1016/j.yofte.2021.102704

6. Xiong-yao X, Cheng-min Y, Wei-ping L, et al. Safety analysis of settlement monitoring data of joints of Yongjiang immersed tube tunnel during operation period. *Chin J Geotechnical Eng* (2019) 41(12):7. doi:10.11779/CJGE201912020

7. Xiong-yao X, Wang P, Yong-sheng L, et al. Monitoring data and finite element analysis of long term settlement of Yongjiang immersed tunnel. *Rock Soil Mech* (2014) 35(8):11. doi:10.16285/j.rsm.2014.08.026

8. Zheng Y, Jinbo Y, Junwei L. Horizontal displacement monitoring technology of the Ningbo Changhong immersed tube tunnel. *Urban Geotechnical Invest Surv* (2014) 35(8):11. doi:10.3969/j.issn.1672-8262.2012.05.044

9. Pan S, Chen C, He J, et al. Design and application of automatic deformation monitoring system (ADMS) for immersed tunnel construction. In: Proceedings of the 2018 Annual Academic Conference of the Chinese Society of Civil Engineering (2018).

10. Liu D, Cui Y, Tan W, Chen Y. Sg-net: Spatial granularity network for one-stage video instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 20-25 June 2021; Nashville, TN, USA (2021). p. 9816–25.

11. Liu D, Cui Y, Chen Y, Zhang J, Fan B. Video object detection for autonomous driving: Motion-aid feature calibration. *Neurocomputing* (2020) 409:1–11. doi:10.1016/j.neucom.2020.05.027

12. Cui Y, Yan L, Cao Z, Liu D. Tf-blender: Temporal feature blender for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 10-17 October 2021; Montreal, QC, Canada (2021). p. 8138–47.

13. Zhang J, Bo Y, Yanfeng G, et al. Research status and prospects of intelligent monitoring technology and information management system for tunnel engineering. *Chin J Underground Space Eng* (2021) 17(02):567–79.

14. Zhao X, Li Q. A review on measurement technology for structural testing in civil engineering. *J Xi'an Univ Arch Tech (Natural Sci Edition)* (2017) 49(01):48–55. doi:10.15986/j.1006-7930.2017.01.008

15. Deng H. *Measurement of vibration response of flexible structure under thermal environment based on machine vision*. Bangladesh: Southeast University (2020).

16. Wu C. *The research of Digital Close-Range Photogrammetry in building deformation monitoring*. Zhengzhou, Henan, China: Henan University of Technology (2009).

17. Wang X. *Development and application of structural displacement monitoring system based on image recognition technology*. China: Nanjing University of Science and Technology (2009).

18. Ye X-W, Chuan-zhi D. Review of computer vision—Based structural displacement monitoring. *China J Highw Transport* (2019) 32(11):19.

19. Chuan-zhi D. *Machine vision-based bridge health monitoring and condition assessment*. Zhejiang, China: Zhejiang University (2016).

20. Tu W, Li Q, Gao W, et al. A real-time precision measurement method for bridge deflection based on machine vision. *Surv Mapp Geographical Inf* (2020) 45(6):80–7. doi:10.14188/j.2095-6045.2020541

21. Li S. Application of close-range photogrammetry in water conservancy and hydropower engineering. *Water Conservancy Hydropower Techn* (1987) 1987(05):19–22.

22. Li H. *Close-range digital imagine technologies for hydroelectric engineering*. China: Hohai University (2005).

23. He L, Zhong L, Qi-jun H, et al. Evaluation for the foundation pit stability based on the multi - sample capacity close - range photogrammetry. *J Saf Environ* (2020) 20(06):2180–6. doi:10.13637/j.issn.1009-6094.2019.1110

24. Meng L, Zou J, Zhu Y. Application of close-range photogrammetry in foundation pit monitoring. *Bull Surv Mapp* (2015) 2015:167–70. doi:10.13474/j.cnki.11-2246.2015.0648

25. Liu C. *Study of photographs monitoring analysis system for slope stability*. Wuhan, China: Wuhan University of Technology (2008).

26. Ling J, Zhang Y, Li M. Research progress of intelligent monitoring system for highway slope. *J Cent South University(Science Technology)* (2021) 52(7):2118–36. doi:10.11817/j.issn.1672-7207.2021.07.003

27. Jin J. *Three-dimensional measurement technology research based on multiple view geometry*. Hefei, China: University of Science and Technology of China (2014).

28. Spencer BF, Jr., Hoskere V, Narazaki Y. Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering* (2019) 5(02):199–222. doi:10.1016/j.eng.2018.11.030

29. Zheng C, Shi H, Chong X. A new monitoring method for large span underground space structure displacement. Proceedings of the 3rd National Engineering Safety and Protection Academic Conference. 2012.

30. 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* (2022) 206:112346. doi:10.1016/j.measurement.2022.112346

31. 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

32. Zhao L, Yang R, Guo C. Research on sub-pixel precise thresholding for feature points of circular array target extraction. *China Sciencepaper* (2015) 10(08):942–7. doi:10.3969/j.issn.2095-2783.2015.08.014

33. Xiangnan M, H L, Lili L. Improved least square algorithm and application in ellipse fitting. Mechatronics engineering school. *Henan Univ Science&Technology* (2014) 35(03):18–21+5. doi:10.15926/j.cnki.issn1672-6871.2014.03.018

34. Tong Z. *Algorithm of relative pose estimation for space target based on monocular-vision*. Harbin, China: Harbin Institute of Technology (2012).

35. Cao Z, Chu Z, Liu D, Chen Y. A vector-based representation to enhance head pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 03-08 January 2021; Waikoloa, HI, USA (2021). p. 1188–97.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 06 March 2023
DOI 10.3389/fphy.2023.1147031

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Kunpeng Wang,
Southwest University of Science and Technology, China
Yu Liu,
Hefei University of Technology, China
Baisen Cong,
Danaher, China

*CORRESPONDENCE
Meiyong Huang,
s200301006@stu.cqupt.edu.cn

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 18 January 2023
ACCEPTED 15 February 2023
PUBLISHED 06 March 2023

CITATION
Chen Y, Huang M, Liu H, Shao K and Zhang J (2023), Real-world low-light image enhancement *via* domain-gap aware framework and reverse domain-distance guided strategy. *Front. Phys.* 11:1147031. doi: 10.3389/fphy.2023.1147031

Real-world low-light image enhancement *via* domain-gap aware framework and reverse domain-distance guided strategy

Yong Chen[1], Meiyong Huang[1]*, Huanlin Liu[2], Kaixin Shao[1] and Jinliang Zhang[1]

[1]The Key Laboratory of Industrial Internet of Things and Network Control, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

Low-light image enhancement (LLIE) has high practical value and development potential in real scenarios. However, the current LLIE methods reveal inferior generalization competence to real-world low-light (LL) conditions of poor visibility. We can attribute this phenomenon to the severe domain bias between the synthetic LL domain and the real-world LL domain. In this article, we put forward the Domain-Gap Aware Framework, a novel two-stage framework for real-world LLIE, which is the pioneering work to introduce domain adaptation into the LLIE. To be more specific, in the first stage, to eliminate the domain bias lying between the existing synthetic LL domain and the real-world LL domain, this work leverages the source domain images *via* adversarial training. By doing so, we can align the distribution of the synthetic LL domain to the real-world LL domain. In the second stage, we put forward the Reverse Domain-Distance Guided (RDDG) strategy, which takes full advantage of the domain-distance map obtained in the first stage and guides the network to be more attentive to the regions that are not compliance with the distribution of the real world. This strategy makes the network robust for input LL images, some areas of which may have large relative domain distances to the real world. Numerous experiments have demonstrated the efficacy and generalization capacity of the proposed method. We sincerely hope this analysis can boost the development of low-light domain research in different fields.

KEYWORDS

real-world low-light image enhancement, domain-gap aware framework, domain adaptation, reverse domain-distance guided strategy, adversarial training

1 Introduction

Real-world LLIE aims to reconstruct normal light images from observations acquired under low-light conditions with low visibility and poor details. Numerous scientific deep-learning approaches [1–5] with the advantage of the powerful capability to learn features [6–10] have been extensively proposed. For the efficient LLIE task [11–13], they recover the visibility and precise details of low-illumination images by learning the relationship between LL and NL images. As the efficiency of deep learning methods is subordinate to the dataset, some methods collect bursts of images with multiple exposure levels captured in real scenarios for real-world LLIE applications [14, 15]. However, since the collections of large-scale paired datasets are incredibly laborious and expensive [16], the existing paired datasets

are usually of small scale, which may cause overfitting when training networks using them. Therefore, some methods have been put forward to enlarge the scale of datasets by synthesizing low-illumination images and forming paired datasets with normal illumination images [17, 18]. However, the synthetic LL images are usually not compliant with real-world distribution, leading to poor generalization capability to the real world for the LLIE methods trained on these datasets [19]. Specifically, the illumination level cannot be improved sufficiently to recover details, or the white balance cannot be maintained correctly. Therefore, it is a worthwhile but challenging task to generate enhancement results that match real-world distribution.

Unsupervised methods are of high practical value and development potential because they do not require paired datasets captured in the same static scenarios [20, 21]. They implement LLIE tasks by taking full advantage of unpaired real-world NL images and LL images. To realize the concept of aligning the distribution of enhanced NL images to the unpaired NL domain, existing methods usually adopt adversarial training directly for the enhanced results against the real-world NL images. Further, to ensure that all regions in the enhanced images are close to the real ones, EnlightenGAN [22] crops image patches randomly from the enhanced images and adopts adversarial training against the real NL image patches. However, these methods seldom notice that the severe domain gap may impede the enhancement performance but only focus on the enhancement procedure, which degrades the generalization performance of the networks trained on synthetic datasets. Moreover, randomly comparing image patches does not guarantee that all regions of the enhanced images match the real-world distribution.

Over recent years, researchers have extensively proposed ways to address the shortage of data with labels for training. For Domain adaptation (DA) methods, the labeled data enables adequate training in the source domain as well as performing new tasks on the unlabeled target domain with new distribution [23]. It greatly improves the effectiveness of methods on the target domain, which is appropriate for real-world LLIE tasks.

In this article, by comprehensively reviewing the potential and reaping the full benefits of alternative methods, we put forward a two-stage framework with the merit of both adversarial learning and domain adaptive methods. Specifically, we propose the Domain-Gap Aware Framework to implement real-world LLIE tasks, which addresses the issue that the input LL images deviate from the real-world distribution.

As shown in Figure 1, the noticeable domain gap between the real-world and the synthetic LL domain can be observed. Besides, different areas in a single low-light image may have different relative domain distances. We find that the domain gap severely degrades the generalization competency of the network to real-world low-light conditions. Therefore, unlike existing methods that ignore the domain bias for synthetic LL images, we propose the Domain-Gap Aware Framework. Specifically, in the first stage, we impose adversarial training on the Darkening Network to eliminate the severe domain gap and generate realistic pseudo-LL images. By doing so, we obtain pseudo-LL images that are consistent with real-world distribution, as well as domain-distance maps. In the second stage, we propose the Reverse Domain-Distance Guided strategy to capitalize fully on the domain-distance maps and mitigate the unrealistic areas of pseudo-LL images. In detail, we assign higher

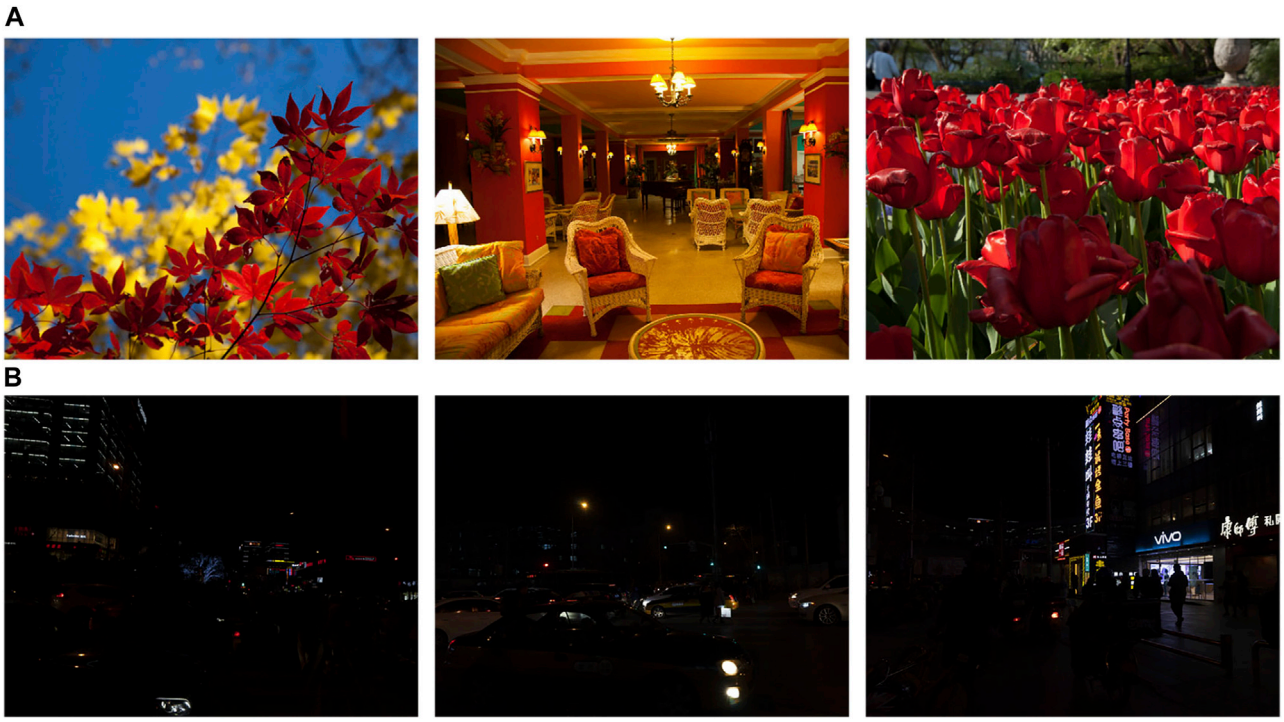


FIGURE 1
(A) The presentation of the existing synthetic LL dataset and **(B)** the real-world LL images. **(A,B)** shows the apparent domain gap in terms of illumination level and white balance lying between the real-world and the existing synthetic LL domain.

weights to the regions in the generated NL images that are relatively far from the real-world domain; while assigning smaller weights to the realistic regions in the training phase, thus mitigating the uncompetitive enhancement competence to real-world scenarios due to the unrealistic input LL patches. The proposed two-stage framework generalizes well to the real world with boosted illumination level and clearly reconstructed structural details, which can significantly facilitate subsequent computer vision tasks and systems [24] focusing on objects at nighttime.

The following are the key contributions to this article:

- We put forward the Domain-Gap Aware Framework to address the domain-gap issue and generate pseudo-low-light images consistent with real-world distribution, which is essential to attain models with high generalization capability for real-world LLIE.
- A Reverse Domain-Distance Guided strategy is proposed for real-world applications. The pixel-wise domain distance maps are taken full advantage of to further promote the robustness of the Enlightening Network. It is worth pointing out that this is the pioneering work to introduce DA to LLIE as far as we know.

The remainder of this paper is structured as follows. In Section 2, we present a brief review of some related works in the LLIE field. Section 3 introduces the proposed framework and strategy. Section 4 shows experimental results to demonstrate the effect of our method, and Section 5 gives a conclusion of the paper.

2 Related work

2.1 CNN-based approaches

CNN-based approaches have become a principal method in the LLIE field with their high efficiency in image analysis [25, 26]. They reconstruct the contrast and structural details of LL images by learning the mapping relationship between LL-NL images. Some methods have collected paired data in real scenarios [27]. However, it is difficult to construct large-scale paired datasets due to the required high cost and heavy workforce. Since the applications of deep learning methods are usually hampered by shortages of data in pairs for training [28], some methods have also made attempts to construct simulated datasets [17, 18, 29]. It is widely known that the data for training are essential for the networks' performance [30]. However, the synthetic dataset was generated under the assumption of simple degradation in terms of illumination level, noise, etc., which leads to the poor generalization of the trained networks to the real world and the side effect, e.g., color distortion and insufficiently improved illumination.

Real-world LLIE has attracted significant research due to its high practical value. Researchers have been extensively designing diverse architectures to achieve better generalization to the real world. In EnlightenGAN [22], the design philosophy is to address the domain gap issue by applying adversarial training using unpaired datasets. In addition, researchers have also made efforts to zero-shot LLIE. Zero-DCE [31] regards the LLIE as a task of curve estimation for image-wise dynamic range adjustment. However, it pays little attention to the domain gap between the to-be-enhanced LL images and the real-world ones and only focuses on the enhanced NL images, which degrades the generalization performance of the networks trained on synthetic datasets.

2.2 Domain adaptation

Domain adaptation (DA) intends to enhance performance when confronting a new target domain despite domain bias [32]. It is beneficial to deal with data shortages for tasks that are difficult to obtain real data.

In this work, we concentrate on eliminating the domain gap to synthesize realistic LL images, which is a preparation phase for the enlightening stage. Inspired by relevant studies in super-resolution [33], we construct a Domain-Distance Aware framework to perform the real-world LLIE. We apply DA to improve the performance of LLIE on real data.

In the next sections, we introduce the proposed Domain-Distance Aware framework and Reverse Domain-Distance Guided strategy in detail.

3 Methods

3.1 Network architecture

Given two domains, which can be described as the LL domain and the NL domain, our goal is to learn an Enlightening Network to promote the visibility and reconstruct structural details of the images in the LL domain while generating enhanced NL estimations belonging to the real-world NL domain. To achieve this objective, we propose the Domain-Gap Aware Framework. We did not follow previous work that directly utilizes the existing synthetic low-illumination datasets to train the Enlightening Network. Instead, our framework takes the domain bias between x^g and x^r into full account. As shown in Figure 2, during the first phase, we train the Darkening Network using adversarial training, which generates pseudo-LL images belonging to the real-world LL domain as well as domain distance maps. Then, in the second stage, we put forward the Reverse Domain-Distance Guided strategy, which leverages the pseudo-LL-NL image pairs and domain distance maps to train the Enlightening Network.

In the next subsection, we first describe how to train the Darkening Network to generate LL-NL image pairs in line with real-world distributions. Then, we show the Reverse Domain-Distance Guided strategy.

3.1.1 Training of darkening network

The general procedure of synthesizing low-light images by existing methods is manually adjusting the illumination and adding noise [17, 18]. However, the illumination levels in the real world are diverse and may also vary spatially in a single image. Moreover, it is difficult to represent noise with simple and known distribution. In a word, the degradations assumed by existing methods are too simple to fully simulate the complex degradation in the real world, which unfortunately leads to domain bias lying between the synthesized LL images and the real-world ones. In contrast, our approach employs a deep

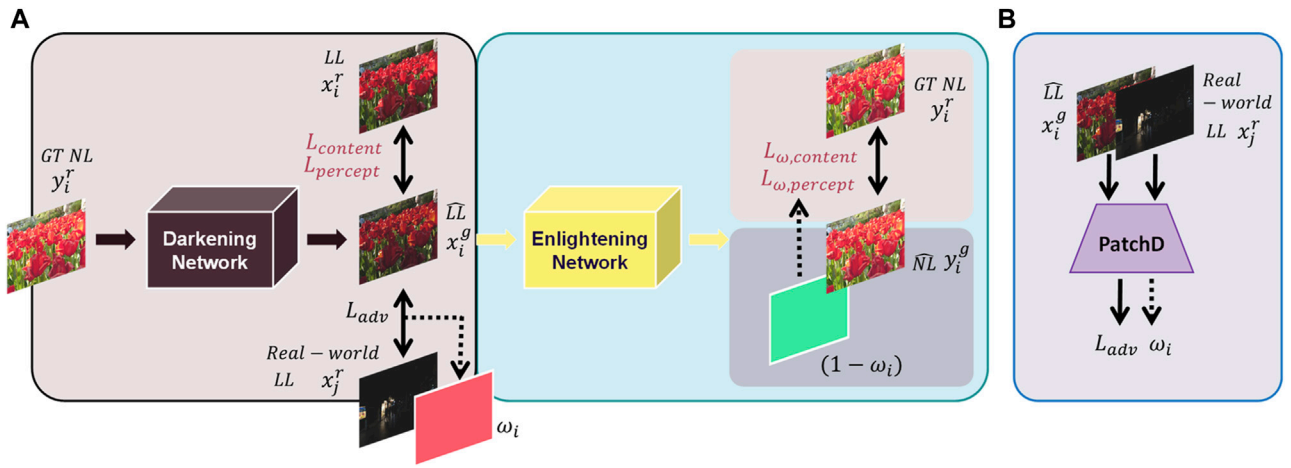


FIGURE 2
(A) The proposed Domain-Gap Aware Framework contains two stages, i.e., the Darkening stage and Enlightening stage. **(B)** Adversarial training during the training of Darkening Network.

network (i.e., the Darkening Network) to learn the real-world degradation process. It works as the generator in the whole framework and extracts the features of NL images using eight blocks (each layer is convolved by a 3 × 3 kernel and activated by a ReLU activation in between).

3.1.1.1 Losses

We employ multiple loss functions to train the Darkening Network. To ensure that the content of the pseudo-LL images is preserved consistently with the GT (Ground-Truth) LL images, we adopt content loss along with the perceptual loss to optimize the distance between them at the image level and the feature level, respectively. In detail, the content loss contains reconstruction loss, which is L1-norm and SSIM (Structural SIMilarity Index) [34] loss, which aims at measuring structural similarities between two images. The reason why we adopt L1-norm as our reconstruction loss is that it treats all errors equally so that the training can keep going even though the error is tiny. Perceptual loss is also widely used in the image reconstruction field, which measures the distance between features extracted *via* deep neural networks.

$$
\begin{aligned}
\boldsymbol{L_{Recons.}} &= \mathbf{E}_{\boldsymbol{y^r}} \left\| \boldsymbol{x_i^r} - \boldsymbol{DN}\left(\boldsymbol{y_i^r}\right) \right\|_1 \\
\mathbf{SSIM}\left(\boldsymbol{m}, \boldsymbol{n}\right) &= \frac{\left(2\boldsymbol{\mu_m}\boldsymbol{\mu_n} + \boldsymbol{c_1}\right)\left(2\boldsymbol{\sigma_{mn}} + \boldsymbol{c_2}\right)}{\left(\boldsymbol{\mu_m^2} + \boldsymbol{\mu_n^2} + \boldsymbol{c_1}\right)\left(\boldsymbol{\sigma_m^2} + \boldsymbol{\sigma_n^2} + \boldsymbol{c_2}\right)} \\
\boldsymbol{L_{SSIM}} &= \mathbf{1} - \mathbf{SSIM}\left(\boldsymbol{x_i^r}, \boldsymbol{DN}\left(\boldsymbol{y_i^r}\right)\right) \\
\boldsymbol{L_{content}} &= \boldsymbol{L_{Recons.}} + \boldsymbol{L_{SSIM}} \\
\boldsymbol{L_{percept}} &= \boldsymbol{E_{y^r}} \left\| \boldsymbol{\Phi}\left(\boldsymbol{x_i^r}\right) - \boldsymbol{\Phi}\left(\boldsymbol{DN}\left(\boldsymbol{y_i^r}\right)\right) \right\|_1
\end{aligned}
$$

We show the adopted loss functions above, where Φ(·) denotes the convolutional layers of the conv5_3 of VGG-16 [35], and SSIM(·,·) means the SSIM score between two input images.

In addition to the above training, to address the domain gap issue and align the distribution of the pseudo-LL images to the real world, the pseudo-LL images are trained against the real-world LL images by adversarial training. Specifically, we adopt a similar strategy as DASR [33], which uses a patch discriminator with four layers of fully convolutional layers to determine whether each image block matches the real-world distribution. This strategy facilitates pseudo-LL images to fit the real-world distribution.

$$
\begin{aligned}
\boldsymbol{L_{Adv.}^G} &= -\boldsymbol{E_{y^r}}\left[\boldsymbol{log D}\left(\boldsymbol{DN}\left(\boldsymbol{y^r}\right)\right)\right] \\
\boldsymbol{L_{Adv.}^D} &= -\boldsymbol{E_{y^r}}\left[\boldsymbol{log D}\left(\boldsymbol{x^r}\right)\right] - \boldsymbol{E_{y^r}}\left[\boldsymbol{log}\left(\mathbf{1} - \boldsymbol{D}\left(\boldsymbol{DN}\left(\boldsymbol{y^r}\right)\right)\right)\right]
\end{aligned}
$$

The loss functions are shown above, where D(·) denotes the patch discriminator.

3.1.2 Reverse domain-distance guided strategy

As shown in Figure 1 previously, each region in the generated LL image may distant diversely from the domain of the real world, i.e., some regions lie relatively close to the domain of the real world, while some regions are relatively far. Since the regions relatively far from the domain of the real world may degrade the enhancement competency of the network, we should endow different regions with diverse attention. We realize this concept by reversing the domain distance maps first and then applying them to eliminate the discrepancy between y_i^g and x_i^r. Thereby, we adaptively adjust the loss functions by assigning diverse weight parameters to these regions adaptively. We present the Reverse Domain-Distance Guided strategy in Figure 3.

3.1.2.1 Losses

We denote the supervised losses as follows, where ω_i denotes the domain distance map for x_i^g, which is attained by the patch Discriminator trained in the first stage.

$$
\begin{aligned}
\boldsymbol{L_{\omega,content}} &= -\boldsymbol{E_{x_i^g, y_i^r}} \left\| \left(\mathbf{1} - \boldsymbol{\omega_i}\right) \odot \left(\boldsymbol{EN}\left(\boldsymbol{x_i^g}\right) - \boldsymbol{y_i^r}\right) \right\|_1 \\
\boldsymbol{L_{\omega,percept}} &= -\boldsymbol{E_{x_i^g, y_i^r}} \left\| \left(\mathbf{1} - \boldsymbol{\omega_i}\right) \odot \left(\boldsymbol{\Phi}\left(\boldsymbol{EN}\left(\boldsymbol{x_i^g}\right)\right) - \boldsymbol{\Phi}\left(\boldsymbol{y_i^r}\right)\right) \right\|_1
\end{aligned}
$$

The trained patch Discriminator can differentiate between the pseudo-LL patches and those from the real-world domain. A smaller value in ω_i means a lower probability that the pseudo-LL patches

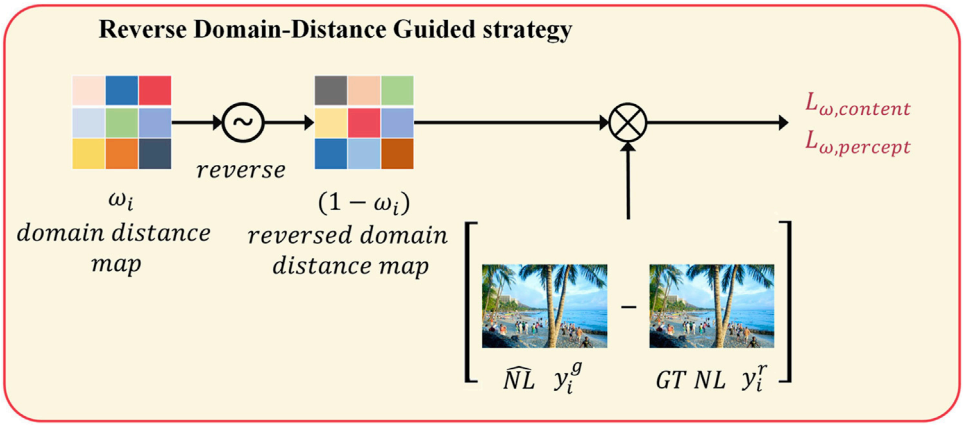


FIGURE 3
The proposed Reverse Domain-Distance Guided strategy. It facilitates the Enlightening Network to be more attentive to the less realistic regions of the input LL images.

belong to the real-world domain. It also indicates a higher value in the reverse of ω_i, i.e., 1- ω_i, and a larger domain distance from the pseudo-LL to the real-world domain. Therefore, we guide the network to be attentive to the enhanced outcomes of the input pseudo-LL patches relatively far from the real-world domain by endowing distance-related importance to different areas. The Reverse Domain-Distance Guided strategy makes full use of domain distance to remedy the unrealistic areas of pseudo-LL images and further improves the generalization to the real world.

To evaluate the proposed method, we describe experimental settings and results in detail in the next section.

4 Experiments

Since the similarity with the ground-truth NL images can reflect the enhanced result to a large extent, we adopt PSNR and SSIM [34] as reference metrics, two widely adopted quality metrics in the image restoration field. In addition, as our method adopts a generative-adversarial network, we also focus on perceptual quality. Therefore, we also adopt LPIPS (Learned Perceptual Image Patch Similarity [36] as the quality metric. Diverse ablation studies are carried out by us to figure out the effect of the proposed strategies in our framework. Then, to figure out our method's generalization competency, the real-world LL dataset is assigned as the testing set. Finally, we further make comparisons with other competing LLIE approaches by applying them to real-world LL datasets.

4.1 Training settings

Researchers have constructed MIT-Adobe FiveK dataset [37], which consists of 5,000 photos retouched by five experts, to adjust the global tone. It has been widely leveraged in the LLIE field. We applied GladNet [38] to the normal-light images retouched by MIT-Adobe FiveK dataset's Expert E to obtain synthetic LL images. We separate 4,000 paired NL-LL images from it to prepare for training and 1,000 paired images to prepare for validation. Then we resize the images to 600 × 400 resolution. Besides, we adopt the DARK FACE dataset (4,000 for training, 1,000 for validation) [39], which consists of 6,000 images obtained under real-world nighttime conditions, as the real-world low-light references.

Let us now turn our attention to the main framework. The network is assigned random weights initially. The Adam method (with momentum and weight decay set to be 0.9 and 0.001, respectively) is adopted to update the network's parameters. Besides, the learning rate is assigned to be 0.0001 initially and then is halved every ten epochs. During the whole training procedure, we maintain a batch size of 16. We carry out all the evaluation experiments on the NVIDIA GeForce GTX3090 and NVIDIA GeForce GTX1080Ti with PyTorch.

4.2 Ablation studies

Before conducting comparison experiments with recently competing methods, we carried out a variety of ablation experiments to delve into diverse loss functions as well as the proposed framework.

4.2.1 Effect of loss functions

We carry out a variety of trying outs for diverse loss functions and figure out the quantitative outcomes on the widely adopted metrics, i.e., PSNR and SSIM, along with LPIPS. During the computation of LPIPS, we extract features of input images through AlexNet [40] to calculate the distance between them. A small LPIPS value means a high similarity. Table 1 displays the quantitative outcomes.

Firstly, let us analyze the effect of each loss function. We can find from the 3rd, 4th, 5th row that in comparison to being supervised by the reconstruction loss and adversarial loss, adding SSIM loss boosts the performance on PSNR, SSIM, and LPIPS metrics with 3.871dB, 0.0654dB, and 0.1177dB, respectively, and adding perceptual loss improves the performance on the three metrics with 3.19dB,

TABLE 1 Ablation study of diverse loss functions and corresponding image quality evaluations.

Recons.	SSIM	Adv	Percept	PSNR↑	SSIM↑	LPIPS↓
√	×	×	×	**20.107**	0.7704	0.1023
√	√	×	×	20.028	**0.7772**	**0.0960**
√	×	√	×	15.761	0.6995	0.2218
√	√	√	×	19.632	0.7649	0.1041
√	×	√	√	18.951	0.7190	0.1880
√	√	√	√	20.079	0.7718	0.1035

Recons., Adv., and Percept. indicate reconstruction loss function, adversarial loss, and perceptual loss, respectively. Scores marked in bold indicate the highest scores on the corresponding metric.

0.0195dB, 0.0338dB, respectively. It indicates the effectiveness of both SSIM loss and perceptual loss in reconstructing texture and details of contents.

Secondly, we can find from the 1st and 2nd row that the best performance is achieved under the settings of using merely content loss, which includes reconstruction loss and SSIM loss. Note that content loss aims at reducing the distance between input images. Therefore, training with them equals supervised learning, which easily achieves better performance in comparison to adversarial training. Nevertheless, our method, which contains adversarial learning for fitting with real-world low-light image distributions, achieves similar quantitative results with supervised learning. In specific, our method performs second best on PSNR and SSIM and obtains rank third on LPIPS with a difference of only 0.028dB, 0.0054dB, and 0.0075 dB with the rank first scores.

As shown qualitatively in Figure 4, our method generates images with sufficiently low exposure levels and correct white balance. Therefore, we can conclude that the LL images generated by our method keep contents consistent with LL images from the existing dataset but closer to the ones captured under poor light conditions. We finally chose $\omega_{col} = 1$, $\omega_{ssim} = 1$, $\omega_{per} = 0.02$, and $\omega_{adv} = 0.02$ for the weight parameters of each loss function.

4.2.2 Effect of the darkening network

4.2.2.1 Comparisons of LL images

We adopt a generative adversarial network to generate pseudo-LL images so that they are close to LL images from the existing dataset in terms of contents while in compliance with the distribution of real-world LL images. Figure 4 presents the contrast between the MIT-Adobe FiveK dataset [37] and pseudo-LL images synthesized by our method. We can find in the 2nd row of the panel (A) and (B) in Figure 4 that the white balance of several images in the existing low-light dataset is going wrong, where white areas of the original NL images appear orange in the existing low-light dataset. This may lead to the color shift in the enhanced images enhanced, which is further proven in Figure 5. Besides, the 2nd row of Figure 4B suggests that the illumination level is not low enough to simulate night lighting conditions. In contrast, the proposed Darkening Network maintains the correct white balance and decreases the illumination level sufficiently in LL versions, as displayed in the 3rd row of the panel (A) and (B) in Figure 4, which facilitates the lightening network to generalize better to the real-world low-light condition.

4.2.2.2 Comparisons of enhancement results

Furthermore, the effect of the proposed Darkening Network is investigated in this subsection. We train the Darkening Network both using pseudo-LL-NL pairs $\{\boldsymbol{x}^g, \boldsymbol{y}^r\}$ and existing paired dataset $\{\boldsymbol{x}^r, \boldsymbol{y}^r\}$, and compare the outcomes in terms of quality and quantity. Figure 5 shows qualitative comparisons. As shown in Figure 5A, we can clearly see that the enhanced outcomes of the existing LL dataset suffer from the color shift. This is because the input LL images are of imperfect white balance, as shown in Figure 4 previously. Besides, as shown in Figure 5C, it easily appears over-exposure, which hinders some regions (e.g., regions in the dark color such as hair, ribbon, skin, and so on.) from retaining semantic darkness, unfortunately. This is due to the insufficient illumination level in existing LL images. Moreover, the enhanced results of the backlit image (in the 5th row of Figure 5C) suffer from artifacts severely. In contrast, as shown in Figures 5B, D, the enhanced results of pseudo-LL images are of correct white balance and appropriate exposure level with good preservation of semantic information, as well as much fewer artifacts introduced to backlit images. Therefore, we confirm that the Enlightening Network can produce superior enhancement results collaborated with the Darkening Network, which fully reflects the effect of the Darkening Network.

Next, we display quantitative comparison results in Table 2.

We can clearly find that training with pseudo-LL-NL pairs $\{\boldsymbol{x}^g, \boldsymbol{y}^r\}$ achieves better scores on PSNR, SSIM, and LPIPS, which exceeds the GT LL-NL pairs $\{\boldsymbol{x}^r, \boldsymbol{y}^r\}$ to a great extent. More specifically, training with $\{\boldsymbol{x}^g, \boldsymbol{y}^r\}$ achieves considerably higher scores on PSNR and SSIM than training with $\{\boldsymbol{x}^r, \boldsymbol{y}^r\}$, both with and without the Reverse Domain-Distance Guided strategy. More specifically, as shown in column2 and 4, training with $\{\boldsymbol{x}^g, \boldsymbol{y}^r\}$ exceeds training with $\{\boldsymbol{x}^r, \boldsymbol{y}^r\}$ by 16.252 dB (= 35.276–18.751 dB) on PSNR and by 0.1022 dB (= 0.9696–0.8674 dB) on SSIM. It demonstrates that the synthesized pseudo-LL-NL pairs are more suitable for real-world LLIE than GT LL-NL pairs from the existing dataset. The reason is that the proposed Darkening Network aims to address the domain gap *via* adversarial training so that the pseudo-LL images match the real-world distribution in terms of exposure level, hue, noise, and so on. However, the existing procedure of synthesizing low-light images assumes simple degradations from NL images, which is far from the complex degradations of the real world. Therefore, we confirm that by adequately taking advantage of target domain data during the training process, the proposed Darkening Network makes a significant contribution to the improvement of enhancement quality.

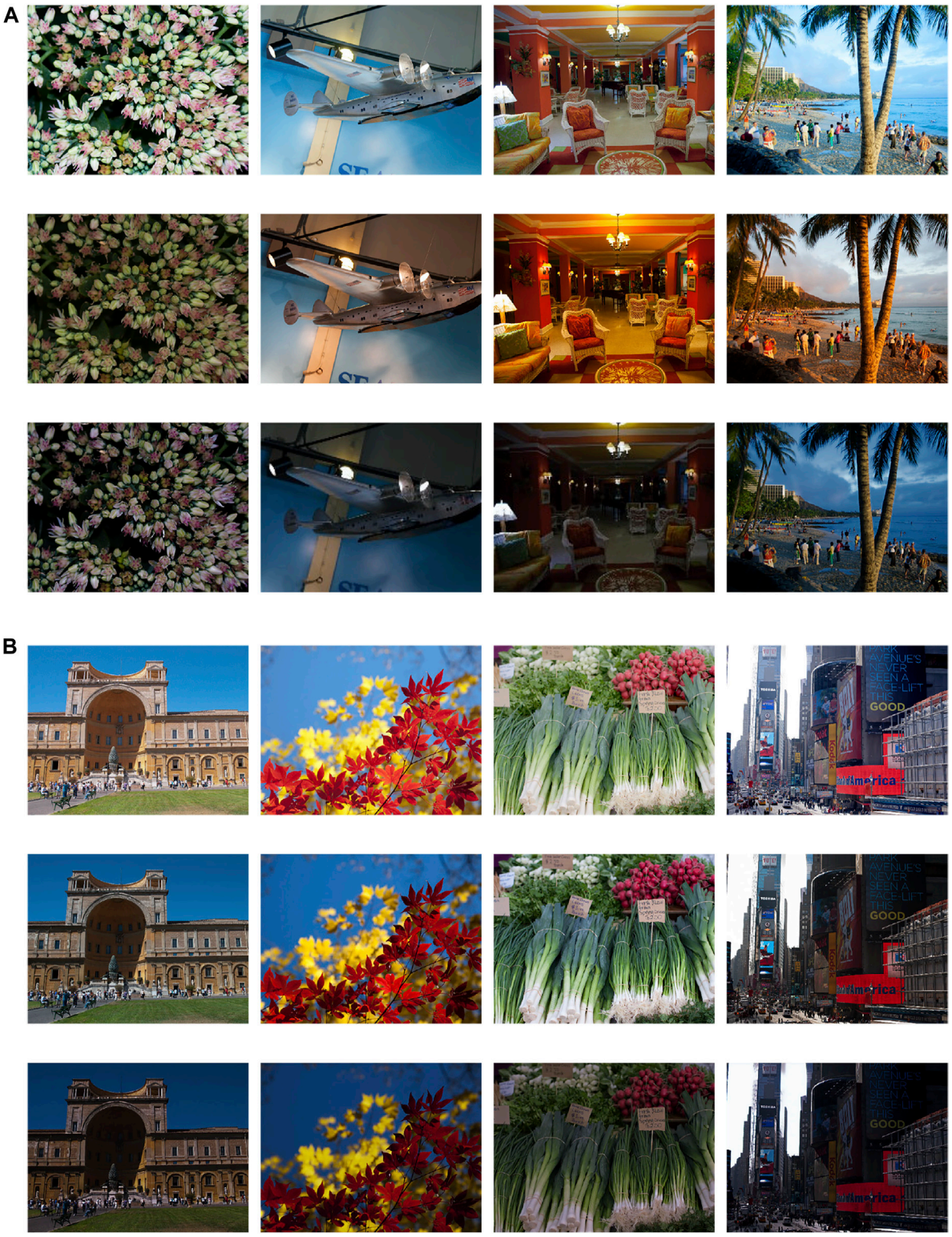


FIGURE 4
The comparison of LL images from the existing dataset and pseudo-LL ones synthesized by our method. **(A)** The comparison of white balance of LL-images; **(B)** The comparison of illumination level. The 1st, 2nd, and 3rd in both **(A)** and **(B)** show original normal-light images, LL images from the existing dataset, and pseudo-LL ones generated by our method, respectively. The images in the 3rd row of each panel are in line with the distribution of the real world in terms of illumination level and white balance.

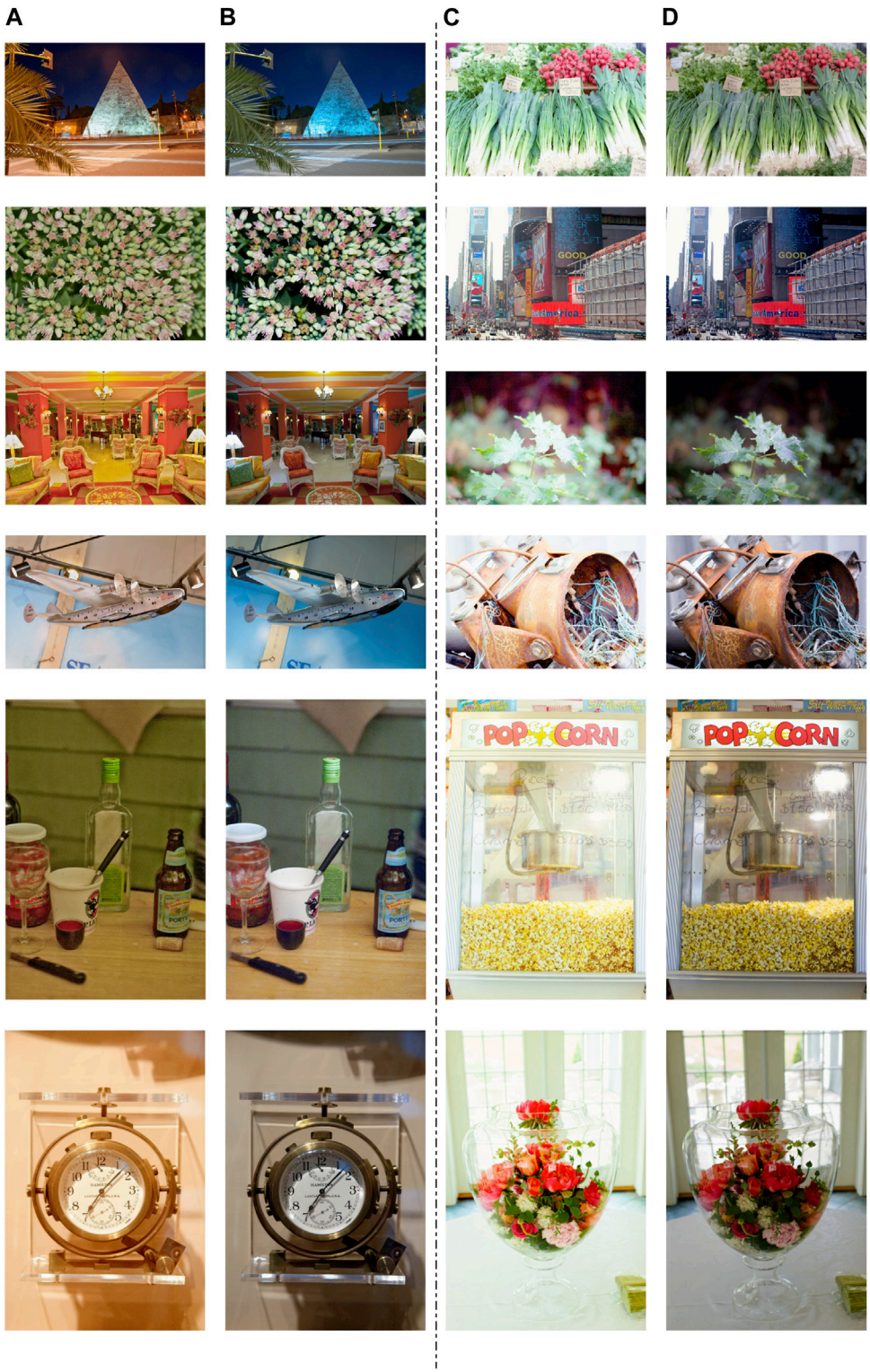


FIGURE 5
Qualitative comparison of the enhancement results of Darkening Network trained with GT LL-NL pairs $\{x^r, y^r\}$ and pseudo-LL-NL pairs $\{x^g, y^r\}$. **(A,C)** shows the enhanced results by GT LL-NL pairs $\{x^r, y^r\}$, **(B,D)** shows the enhanced results by pseudo-LL-NL pairs $\{x^g, y^r\}$. The proposed Darkening Network performs better by retaining semantically dark regions and correct white balance.

TABLE 2 Ablation studies of Darkening Network and Reverse Domain-Distance Guided strategy with the settings of both pseudo-LL-NL pairs and GT LL-NL pairs from the existing dataset.

Reverse domain-distance guided strategy	$\{x^r, y^r\}$		$\{x^g, y^r\}$	
	√	×	√	×
PSNR↑	18.796	18.751	**37.257**	35.276
SSIM↑	0.8618	0.8674	**0.9731**	0.9696
LPIPS↓	0.1764	0.1674	0.1587	**0.1547**

The scores marked in bold indicate the highest scores on the corresponding metric.

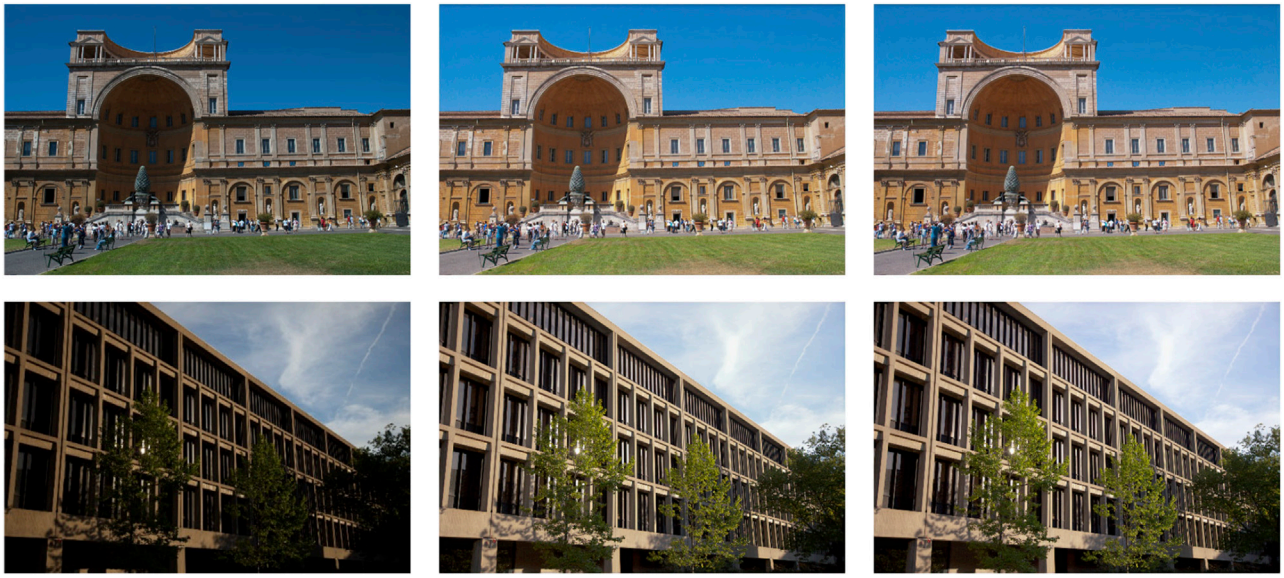


FIGURE 6
Qualitative comparison of the enhancement results without and with Reverse Domain-Distance Guided strategy. The texture and details are better reconstructed with the propsed strategy.

4.2.3 Effect of reverse domain-distance guided strategy

To verify the effectiveness of the Reverse Domain-Distance Guided strategy, we conduct ablation studies with the settings of training with $\{x^g, y^r\}$, and $\{x^r, y^r\}$. Table 2 and Figure 6 show quantitative and qualitative results, respectively. For convenient comparison, the quantitative outcomes are displayed in Table 2. We can easily discover that adopting the Reverse Domain-Distance Guided strategy improves the PSNR and SSIM with a certain magnitude of 1.981 dB (=37.257–35.276 dB) on PSNR and 0.0035 dB (= 0.9731–0.9696 dB) on SSIM when training with $\{x^g, y^r\}$. The reason is that domain distances between pseudo-LL images x^g and real-world LL images x^r are taken full advantage of at the enhancement stage. Specifically, the Enlightening Network is driven to emphasize the regions that are not in compliance with the real world by allocating greater weight to them during the training process. Therefore, it is easy to understand that the collocation of the Reverse Domain-Distance Guided strategy and the proposed Darkening network is beneficial to the reconstruction of texture and details with the pseudo-LL-NL pairs $\{x^g, y^r\}$.

Let us investigate the effect of the Reverse Domain-Distance Guided strategy. We can find from Figure 6 that those semantically dark regions maintain their semantic darkness during the improvement of illumination level without under-exposure for other regions, which facilitates the images to appear more realistic. Therefore, the proposed Reverse Domain-Distance Guided strategy is of significance for the generalization of LLIE to the real world.

4.3 Evaluations of generalization on the real-world dataset

The Exclusively Dark dataset [41] is proposed to facilitate better detection under poor visibility conditions for nighttime systems and applications. It contains a total of 7,363 images of 12 specified object categories. Some images were sub-sampled from existing large-scale datasets, including Microsoft COCO, PASCAL VOC, and ImageNet. We carry out evaluations for the generalization capacity on the Exclusively Dark dataset and DARK FACE dataset. Figure 7 shows a

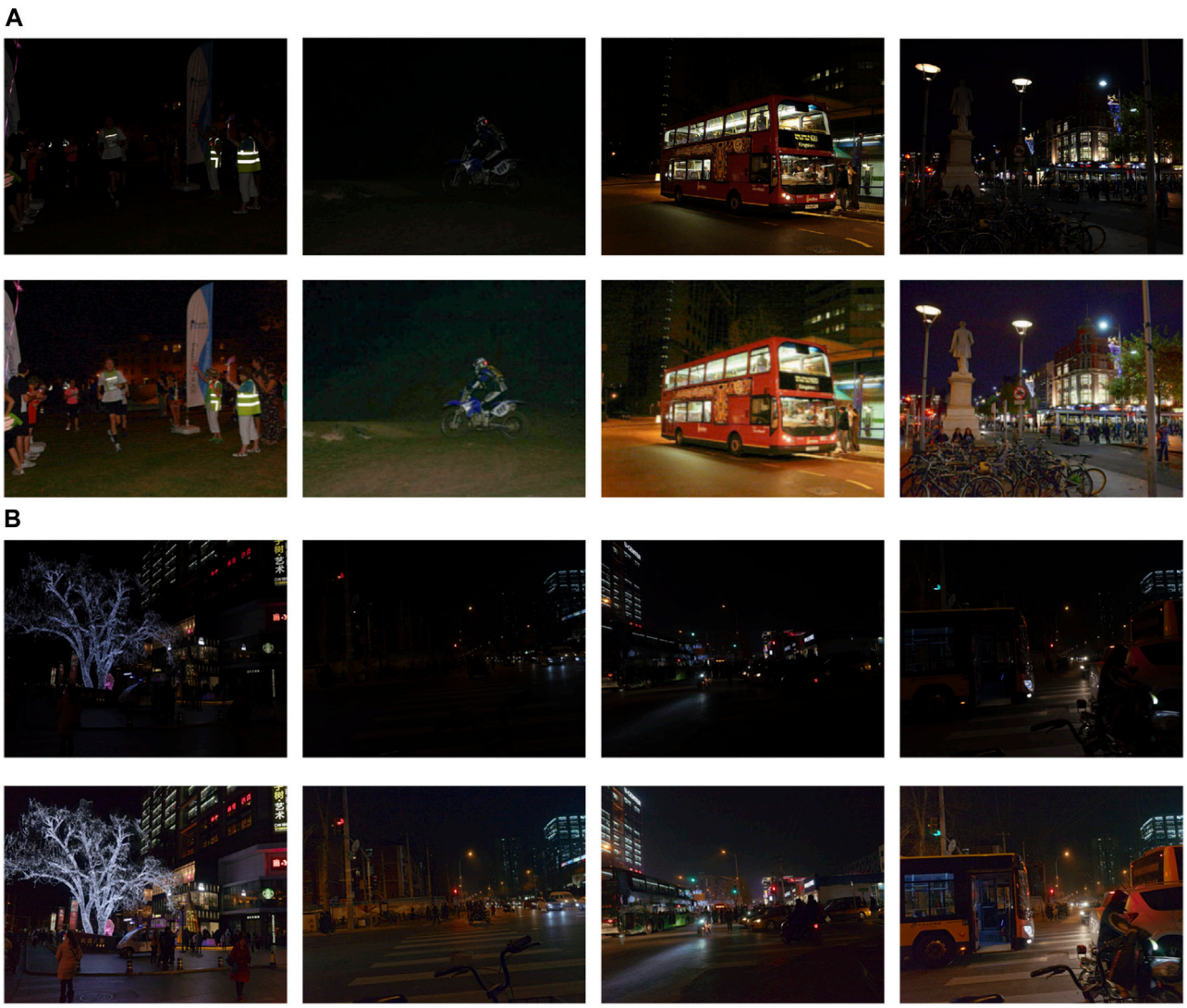


FIGURE 7
Evaluation of generalization capability on real-world datasets, i.e., **(A)** Exclusive Dark dataset and **(B)** DARK FACE dataset. In both **(A,B)** panels, the 1st row indicates input LL images, and the 2nd row shows enhanced results by the proposed framework. Our method has superior generalization capability to extremely low-illumination real-world conditions.

visual representation of enhancement results. Numerous outcomes demonstrate that the proposed approach can greatly boost the visibility of objects under extremely poor or significant variational illumination conditions. Therefore, we can confirm that the proposed approach can generalize well to extremely dark light conditions in the real world. Furthermore, our method can potentially facilitate subsequent computer vision systems for night vision surveillance since the performance of object-focused works usually drops when the given images are degraded [42, 43].

4.4 Comparative experiments with state-of-the-arts

Let us conduct comparative experiments with recent competing LLIE approaches on the DARK FACE dataset [39]. Figure 8 displays the qualitative contrastive results of different competing approaches, including EnlightenGAN [22], DSLR [44], DRBN [45], TBEFN [46], RRDNet [47], and MBLLEN [48]. Qualitative results show that all the methods can effectively enhance the LL images captured under severely real-world low-light nighttime environments in terms of illumination level. However, some methods introduce side effects. Specifically, it can be clearly found that the overall hue of the image is distorted in EnlightenGAN. Besides, DRBN, TBEFN, and MBLLEN introduce distinct artifacts to local areas. It can be concluded that DSLR, RRDNet, and our method attain the top three best performances. Let us further investigate their differences in detail. It can be clearly found that DSLR and RRDNet tend to generate blur artifacts, i.e., structural details cannot be clearly reconstructed. Besides, RRDNet cannot sufficiently improve the illumination level. In contrast, our method improves visibility without introducing blurriness and shows a better reconstruction

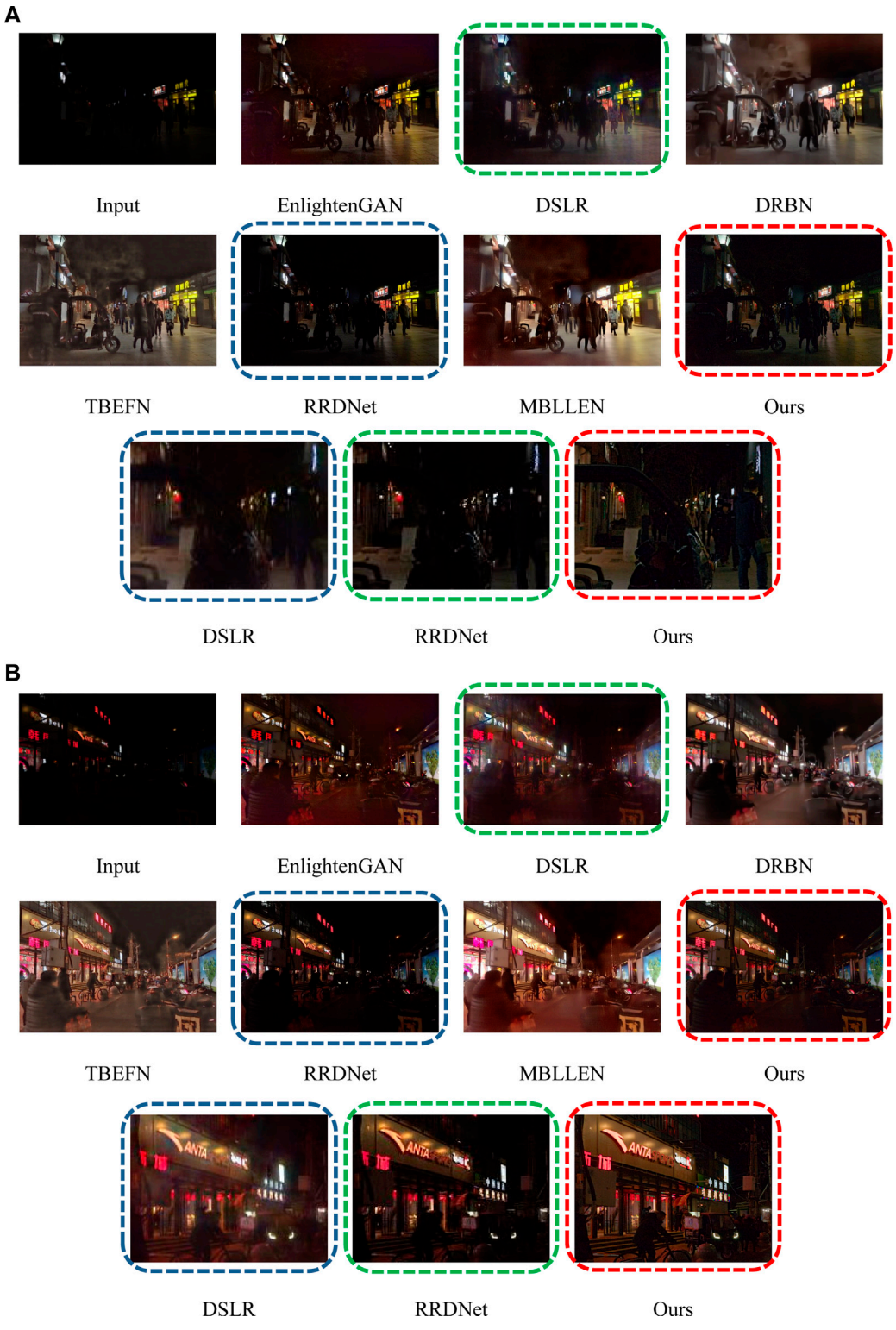


FIGURE 8
Vivid qualitative enhancement outcomes of recently competing network structures and our framework on the DARK FACE dataset. We present the results of SOTA methods on two specified images. We further compare the three most competing methods marked with boxes by zooming in on the local area of their enhanced results. Our method achieves the more superior enhancement results for real-world LL images than SOTA methods in respect of structural details and visibility.

of details, as shown in the zoomed-in comparison results. Therefore, we can confirm that our approach works most effectively relative to other recently competing LLIE methodologies.

Finally, we give a conclusion in Section 5.

5 Conclusion

This paper introduces domain adaptation to the LLIE field. Unlike previous methods that directly adopt existing synthetic low-light datasets, we propose the Domain-Gap Aware Framework, which addresses the dilemma of domain-gap lying between pseudo-LL and real-world LL domain. To eliminate the domain gap, we employ adversarial training to the Darkening Network in the first stage and obtain domain distance maps. In the second stage, we put forward a Reverse Domain-Distance Guided (RDDG) strategy, which further drives the enhancement network to focus on the regions that are not consistent with real-world distribution. In the second stage, we put forward a Reverse Domain-Distance Guided (RDDG) strategy, which further guides the Enlightening network to be attentive to the regions that are not consistent with real-world distribution. We objectively validate the effect of our framework on real-world LL datasets and conduct comparative experiments with other methods. Prominent experimental outcomes present that our framework outperforms other competing network structures.

In our future endeavors, we will explore more contributory approaches for the LLIE field. In addition, we will introduce LLIE methods to subsequent computer vision tasks and systems for diverse applications, such as driving assistant systems and nighttime surveillance.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://flyywh.github.io/CVPRW2019LowLight/, https://github.com/weichen582/GLADNet.

Author contributions

YC and MH contributed to putting forward core conceptions and design of the study; YC provided the computing platform; HL, KS, and JZ organized the database; YC and MH performed the statistical analysis and wrote the first draft of the manuscript; YC contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

The authors are grateful for the provision of computing platforms and academic guidance by Jiaxu Leng.

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. Afifi M, Derpanis KG, Bjorn O, Michael SB. Learning multi-scale photo exposure correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 20 2021 to June 25 2021; Nashville, TN, United States. IEEE (2021). p. 9157–67.

2. Gao X, Lu W, Zha L, Hui Z, Qi T, Jiang J. Quality elevation technique for UHD video and its VLSI solution. *J Chongqing Univ Posts Telecomm: Nat Sci Ed* (2020) 32(5): 681–97. doi:10.3979/j.issn.1673-825X.2020.05.001

3. Setyaningsih E, Wardoyo R, Sari AK. Securing color image transmission using compression-encryption model with dynamic key generator and efficient symmetric key distribution. *Digital Commun Networks* (2020) 6(4):486–503. doi:10.1016/j.dcan.2020.02.001

4. Xu X, Liu H, Li Y, Zhou Y. Image deblurring with blur kernel estimation in RGB channels. *J Chongqing Univ Posts Telecomm: Nat Sci Ed* (2018) 30:216–21. doi:10.3979/j.issn.1673-825X.2018.02.009

5. 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* (2020) 70:1–23. doi:10.1109/TIM.2020.3024335

6. Kyle M, Zhai X, Yu W. Image analysis and machine learning-based malaria assessment system. *Digital Commun Networks* (2022) 8(2):132–42. doi:10.1016/j.dcan.2021.07.011

7. Zhang G, Jian W, Yi Y. Traffic sign recognition based on ensemble convolutional neural network. *J Chongqing Univ Posts Telecomm: Nat Sci Ed* (2019) 31(4):571–7. doi:10.3979/j.issn.1673-825X.2019.04.019

8. 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): 8062–72. doi:10.1109/JSEN.2020.2981719

9. 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

10. 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

11. Yang S, Zhou D, Cao J, LightingNet GY. LightingNet: An integrated learning method for low-light image enhancement. *IEEE Trans Comput Imaging* (2023) 9:29–42. doi:10.1109/TCI.2023.3240087

12. Fan S, Liang W, Ding D, Yu H. Lacn: A lightweight attention-guided ConvNeXt network for low-light image enhancement. *Eng Appl Artif Intel* (2023) 117:105632. doi:10.1016/j.engappai.2022.105632

13. Cotogni M, Cusano C. TreEnhance: A tree search method for low-light image enhancement. *Pattern Recognit* (2023) 136:109249. doi:10.1016/j.patcog.2022.109249

14. Cai J, Gu S, Zhang J. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans Image Process* (2018) 27(4):2049–62. doi:10.1109/TIP.2018.2794218

15. Chen C, Chen Q, Xu J, Koltun V. Learning to see in the dark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR); June 18 2018 to June 22 2018; Salt Lake City, UT, United States. IEEE (2018). p. 3291–300.

16. Li C, Guo C, Han L, Jiang J, Cheng M-M, Gu J, et al. Low-light image and video enhancement using deep learning: A survey. *IEEE Trans Pattern Anal Mach Intell* (2021) 44(12):9396–416. doi:10.1109/TPAMI.2021.3126387

17. Wang L-W, Liu Z-S, Siu W-C, Lun DP. Lightening network for low-light image enhancement. *IEEE Trans Image Process* (2020) 29:7984–96. doi:10.1109/TIP.2020.3008396

18. Lv F, Li Y, Lu F. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *Int J Comput Vis* (2021) 129(7):2175–93. doi:10.1007/s11263-021-01466-8

19. Shi Y, Wu X, Zhu M. *Low-light image enhancement algorithm based on retinex and generative adversarial network* (2019). *arXiv preprint arXiv:1906.06027*.

20. Zhu J, Meng L, Wu W, Choi D, Ni J. Generative adversarial network-based atmospheric scattering model for image dehazing. *Digital Commun Networks* (2021) 7(2):178–86. doi:10.1016/j.dcan.2020.08.003

21. Jin S, Qi N, Zhu Q, Ouyang H. Progressive GAN-based transfer network for low-light image enhancement. In: Proceeding of the 28th International Conference on Multimedia Modeling (MMM); June 6 2022 to June 10 2022; PHU QUOC, Vietnam. IEEE (2022). p. 292–304.

22. Jiang Y, Gong X, Liu D, Cheng Y, Fang C, Shen X, et al. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans Image Process* (2021) 30:2340–9. doi:10.1109/tip.2021.3051462

23. Wang M, Deng W. Deep visual domain adaptation: A survey. *Neurocomputing* (2018) 312:135–53. doi:10.1016/j.neucom.2018.05.083

24. Leng J, Liu Y, Wang Z, Hu H, CrossNet GX. CrossNet: Detecting objects as crosses. *IEEE Trans Multimedia* (2021) 24:861–75. doi:10.1109/TMM.2021.3060278

25. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, et al. Learning enriched features for real image restoration and enhancement. In: Proceeding of the European Conference on Computer Vision (ECCV); February 7 2020 to February 20 2020; Hilton Midtown, New York, United States (2020). p. 492–511.

26. Xu K, Yang X, Yin B, Lau RW. Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR); June 13 2020 to June 19 2020; Seattle, United States. IEEE (2020). p. 2281–90.

27. Wei C, Wang W, Yang W, Liu J. *Deep retinex decomposition for low-light enhancement* (2018). *arXiv preprint arXiv:1808.04560*.

28. Wang Y, Liu D, Jeon H, Chu Z, Matson ET. End-to-end learning approach for autonomous driving: A convolutional neural network model. In: Proceedings of the 15th International Conference on Agents and Artificial Intelligence; February 22 2023 to February 24 2023; Lisbon, Portugal. IEEE (2019). p. 833–9.

29. Liu D, Cui Y, Cao Z, Chen Y. A large-scale simulation dataset: Boost the detection accuracy for special weather conditions. In: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN); July 19 2020 to July 24 2020; Glasgow, United Kingdom. IEEE (2020). p. 1–8.

30. Wang Q, Fang Y, Ravula A, Feng F, Quan X, WebFormer LD. The web-page transformer for structure information extraction. In: Proceedings of the ACM Web Conference 2022; April 25 2022 to April 29 2022; Virtual Event, Lyon France. IEEE (2022). p. 3124–33.

31. Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, et al. Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 13 2020 to June 19 2020; Seattle, United States. IEEE (2020). p. 1780–9.

32. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* (2010) 22(10):1345–59. doi:10.1109/TKDE.2009.191

33. Wei Y, Gu S, Li Y, Timofte R, Jin L, Song H. Unsupervised real-world image super resolution via domain-distance aware training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 20 2021 to June 25 2021; Nashville, TN, United States. IEEE (2021). p. 13385–94.

34. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. *IEEE Trans Image Process* (2004) 13(4):600–12. doi:10.1109/TIP.2003.819861

35. Simonyan K, Zisserman A. *Very deep convolutional networks for large-scale image recognition* (2014). *arXiv preprint arXiv:1409.1556*.

36. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 18 2018 to Jun 22 2018; Salt Lake City, UT, United States. IEEE (2018). p. 586–95.

37. Bychkovsky V, Paris S, Chan E, Durand F. Learning photographic global tonal adjustment with a database of input/output image pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 21 2011 to June 23 2011; Seattle, United States. IEEE (2011). p. 97–104.

38. Wang W, Wei C, Yang W, Liu J. Gladnet: Low-light enhancement network with global awareness. In: Proceedings of the 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018); May 15 2018 to May 19 2018; Xi'an, China. IEEE (2018). p. 751–5.

39. Yuan Y, Yang W, Ren W, Liu J, Scheirer WJ, Wang Z. *UG^{2+} track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments* (2019). *arXiv preprint arXiv:1904.04474*. doi:10.1109/TIP.2020.2981922

40. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM* (2017) 60(6):84–90. doi:10.1145/3065386

41. Loh YP, Chan CS. Getting to know low-light images with the exclusively dark dataset. *Comput Vis Image Underst* (2019) 178:30–42. doi:10.1016/j.cviu.2018.10.010

42. Leng J, Wang Y. RCNet: Recurrent collaboration network guided by facial priors for face super-resolution. In: Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME); July 18 2022 to July 22 2022; Taipei, Taiwan, China. IEEE (2022). p. 1–6.

43. Leng J, Mo M, Zhou Y, Gao C, Li W, Gao X, et al. Pareto refocusing for drone-view object detection. *IEEE Trans Circuits Syst Video Techn* (2022). doi:10.1109/TCSVT.2022.3210207

44. Lim S, Kim W. Dslr: Deep stacked laplacian restorer for low-light image enhancement. *IEEE Trans Multimedia* (2020) 23:4272–84. doi:10.1109/TMM.2020.3039361

45. Yang W, Wang S, Fang Y, Wang Y, Liu J. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE Trans Image Process* (2021) 30:3461–73. doi:10.1109/TIP.2021.3062184

46. Lu K, Zhang L. Tbefn: A two-branch exposure-fusion network for low-light image enhancement. *IEEE Trans Multimedia* (2021) 23:4093–105. doi:10.1109/TMM.2020.3037526

47. Zhu A, Zhang L, Shen Y, Ma Y, Zhao S, Zhou Y. Zero-shot restoration of underexposed images via robust retinex decomposition. In: Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME); July 6 2020 to July 10 2020; London, United Kingdom. IEEE (2020). p. 1–6.

48. Lv F, Lu F, Wu J, Lim C. Mbllen: Low-light image/video enhancement using CNNs. In: Proceedings of the British Machine Vision Conference (BMVC); November 21 2022 to November 24 2022; London, United Kingdom (2018). p. 4.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 08 March 2023
DOI 10.3389/fphy.2023.1159531

OPEN ACCESS

EDITED BY
Bo Xiao,
Imperial College London,
United Kingdom

REVIEWED BY
Guanqiu Qi,
Buffalo State College, United States
Rixing Zhu,
Shijiazhuang Tiedao University, China
Chao Qian,
Chang'an University, China

*CORRESPONDENCE
Hao Ding,
dinghao@cmhk.com
Jianzhong Chen,
chenjianzhong@cmhk.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 06 February 2023
ACCEPTED 22 February 2023
PUBLISHED 08 March 2023

CITATION
Yang M, Lu S, Ding H and Chen J (2023), Traffic safety assessment method of the immersed tunnel based on small target visual recognition image.
Front. Phys. 11:1159531.
doi: 10.3389/fphy.2023.1159531

Traffic safety assessment method of the immersed tunnel based on small target visual recognition image

Meng Yang[1,2], Shanfeng Lu[3], Hao Ding[1,2]* and Jianzhong Chen[1,2]*

[1]China Merchants Chongqing Communications Technology Research and Design Institute Co, Ltd., Chongqing, China, [2]National Engineering Research Center for Highway Tunnel, Chongqing, China, [3]Guangxi Xinhengtong Expressway Co, Ltd., Guangxi, China

The quality of lighting installation performance has a direct impact on the traffic safety of immersed tunnels. To effectively investigate and judge the traffic safety of immersed tunnels having different lighting installations, a traffic safety assessment method for immersed tunnels based on lighting performance degradation was put forward in this study by using big data technology. Numerical simulation was used to simulate the lighting environment in an immersed tunnel under different conditions of lighting performance degradation, conduct the small target recognition test in a physical tunnel, and calculate the traffic safety factor; then, a real-time kinematic assessment model of traffic safety in immersed tunnels was built in combination with the key index factors influencing lighting installations in immersed tunnels. The test results showed that the performance degradation of lighting installations positively correlated with the visual cognition of drivers and passengers. long short-term memory neural network model can effectively assess the traffic safety of immersed tunnels, and the root mean square error (RMSE) and coefficient of determination of the model were separately 1.029 and 0.95, which were superior to the RMSE and coefficient of determination of random forest and recurrent neural network model, and the running time was often less than 1min, complying with the rea; -time assessment requirements; the boundary value of the traffic safety factor of immersed tunnels was 0.6304, and if a value was less than the boundary value, it indicated that the performance of lighting installations was not good and might pose a threat to traffic safety. The research results provided a new perspective for the status assessment of lighting installations in immersed tunnels and also offered a theoretical basis for fine maintenance and repairs of lighting installations.

KEYWORDS

immersed tunnel, deep learning, luminaire failure, traffic safety, safety assessment

1 Introduction

In China, immersed tunnels provide convenience for the production and life of travelers in crossing rivers and deep sea, and the operating safety of immersed tunnels is always the focus of people [1, 2]. In an immersed tunnel, nearly 80% of traffic information is obtained through vision. To provide a lighting environment for immersed tunnels, lighting installations shall uninterruptedly operate 24 h a day, so the quality of their performance has a direct impact on the traffic safety of immersed tunnels [3, 4]. The lighting installation performance is correlated with its components and parts as well as the operating

environment. The ventilation and heat dissipation in an underwater immersed tunnel are limited and the humidity and salinity in it are relatively high [5, 6], causing more serious damage to the electrical parts in a lighting installation.

And meanwhile, pollutants produced by traffic in the tunnel adhere to the surface of lighting installations [7–9], which will also result in the performance degradation of lighting installations and even cause a failure of lighting installations.

The maintenance and repair of immersed tunnels can retard the performance degradation of lighting installations, but the existing research into the maintenance and repairs of tunnel lighting installations and the relevant specifications continue to follow the standard for highway tunnel assessment; the availability of equipment is used as the unique indicator of status assessment to develop the plan of maintenance and repairs [10], and there is a lack of theoretical research into different maintenance contents of lighting installations under different performance conditions. In the immersed tunnel lighting design, the maintenance factor will be determined, but it is easy to cause redundancy in preliminary lighting and inadequacy of later lighting and different immersed tunnels and different types of lighting installations vary greatly [11–13], and the method of established indicators and empirical discrimination cannot accurately determine the performance status of lighting installations and it is difficult to effectively carry out fine maintenance and repairs, bringing a hidden danger to traffic safety of immersed tunnels.

The research into the facility performance degradation is mainly applied to the key fields of science and technology, such as aerospace, nuclear power and large internal combustion engine in the early years [14–16], and as the information technology develops and the data acquisition and mining technology becomes mature, the facility performance degradation has been gradually applied to other fields, including wind power [17, 18], mines, E&M and mechanical engineering [19–22]. XU Zhen et al. utilized the Internet of Things (IoT) technology to collect tunnel E&M equipment information, build an operating state judgment model of tunnel E&M equipment, a monitoring state assessment model of tunnel E&M equipment and a medium-term prediction model for the use of key tunnel equipment, and comprehensively analyze the technical status of tunnel E&M equipment [23]; Zhang et al. and JIN Yinli et al. analyzed and established the weights of influencing factors in different layers according to the structural characteristics and maintenance quality status of the highway E&M system equipment, and used the analytic hierarchy process (AHP) and the fuzzy mathematics theory to comprehensively evaluate the operation of E&M system facilities [24, 25]. Cui et al. analyzed the law of variations in the performance of E&M equipment with the operating life, deduced the hierarchy standard system, calculated the importance of various standards with the Delphi method, and built a fuzzy synthetic evaluation model (FSEM) [26]; ZHU Liwei put forward 4 types of real-time recognition models based on the data transfer path of the highway tunnel E&M system equipment and the topology of corresponding functions for the operating status of functions of the highway tunnel E&M system [27]; ZHANG Jianping et al. put forward a data model for optoelectronic facility luminance attenuation based on Weibull distribution and simulated the law of facility performance degradation with the parameter fitting method [28]. However, the existing monitoring data have not been effectively used, resulting in a waste of monitoring data. At the same time, there is relatively little research on the impact of electromechanical facilities or lighting facilities failure on tunnel traffic safety, and the impact of lighting facilities performance degradation on traffic safety is even blank. Therefore, how to tap the impact of performance degradation of lighting installations on traffic safety of immersed tunnels without influencing normal traffic operation according to the existing monitoring data is the key to the current traffic safety assessment of immersed tunnels.

In this study, the research into the law of variations in the performance degradation of lighting installations and the lighting environment of immersed tunnels was conducted according to numerical simulation and field test data, and based on the results of small target recognition, the traffic safety factor of immersed tunnels was established; based on the key indicators influencing the performance degradation of lighting installations, the LSTM neural network was utilized to build a traffic safety assessment model of immersed tunnels to realize the real-time kinematic assessment of traffic safety of immersed tunnels. The research results can be directly applied to the fine operation and maintenance of the lighting facilities in immersed tunnel, which can fully perceive the advantages and disadvantages of lighting facilities in real time, accurately determine the driving safety under the action of lighting facilities, and ensure the safe operation of the tunnel.

2 Impact of the lighting performance degradation on traffic safety

The performance degradation of lighting installations in an immersed tunnel will change the lighting environment in the immersed tunnel, which will affect the visual discrimination of drivers and passengers. At present, the performance evaluation of lighting installations in immersed tunnels is mainly carried out through regular spot check of pavement luminance from local areas, but due to the influence of harsh environment and exhaust gas pollution, the performance of lighting installations is subject to continuous and kinematic degradation, and the traditional evaluation methods are difficult to effectively judge tunnel pavement luminance and traffic safety. Consequently, the methods of numerical simulation and simulation test are used to simulate the traffic safety of immersed tunnels at different degrees of performance degradation, establish the traffic safety factor of immersed tunnels and realize the traffic safety assessment of immersed tunnels in this paper.

2.1 Simulation of lighting performance degradation

2.1.1 Simulation model building

The model uses DIALux4.13, which is the software for professional lighting design, to calculate and analyze the luminous effect of tunnel pavement under the performance degradation of lighting installations in immersed tunnels; the cross section dimensions of the model are the same as the actual tunnel dimensions, and the tunnel clearance is 7.25 m and the width is 12.75 m, the lane width is 3.75 m, and a 0.75 m overhaul access is

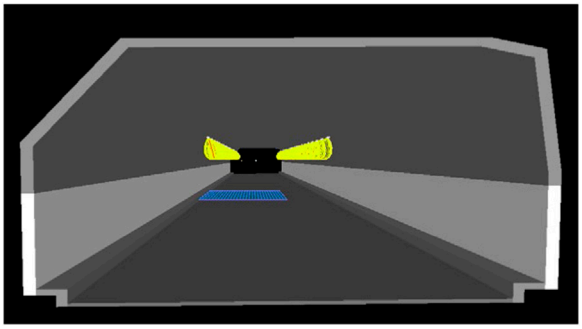


FIGURE 1
Schematic diagram of tunnel simulation.

reserved on each of left and right sides. For the actual engineering pavement, the porous asphalt pavement is designed; the pavement reflection characteristic is set to be R3 (which is predominantly diffuse reflection, with some mirror reflection), the pavement reflectivity is 0.22, the glossiness is S_1 1.1, the wall reflectivity below 2.75 m on both side walls is set to be 0.7, and the average luminance coefficient is Q_0 0.07. The schematic diagram of the tunnel model is shown in Figure 1.

The 60W LED lamps are used for lighting simulation, with the interval of lamps of 4.5 m, The initial luminous flux of LED lamps is 7,200 lm, the correction factor is 0.98, the mounting height of lamps is 5.5 m, the normal angle of the luminous surface of lamps is the same as that of the vertical surface, and the consistent dip angle of lamps is kept.

For the grid in the lighting installation testing area, the computational grid of pavement illuminance and luminance with a longitudinal length of 27 m and a transverse length of 7.5 m is laid, and the grid computation size is 30 × 20. There are 20 transverse calculating points and 30 longitudinal calculating points in the grid.

2.1.2 Simulation results and verification

In DIALux software, the luminous flux of luminaires is set to simulate the results of performance degradation of lighting installations. To verify the accuracy of calculated results of the simulation model, the grid method was adopted to test the lighting environment of the pavement of a physical immersed tunnel in this paper, and the longitudinal length of the testing area was 10m, the transverse length was 7.5 m, and the measuring space was 1 m. The pavement illuminance and uniformity were tested at 100% of luminous flux and the results were compared with the numerical simulation, as shown in Table 1.

It can be seen from the table that the relative error between the measured results of the pavement luminance in the middle section of the immersed tunnel and the average illuminance of simulation results in DIALux software is 0.08%, the relative error of uniformity of luminance is 1.93%, and the relative error of longitudinal uniformity was 2.13%, indicating that there is a small difference between the measured pavement illuminance of immersed tunnels and the results of numerical simulation, and the simulation results are more accurate.

In the simulation model, the results of tunnel pavement luminance at different luminous fluxes (i.e. 100% (without performance degradation), 90%, 80% and 70%) were separately simulated. According to the requirements of LED lighting, the lighting installations will fail if the luminous flux is less than 70% [29], so such simulation is not conducted. The immersed tunnel luminance is divided by colors, and the results of luminance simulation of lighting installations in an immersed tunnel are shown in Figure 2.

It can be seen from Figure 2 that the illuminance of tunnel pavement and overhaul access is significantly decreased with a reduction in the luminous flux of lighting installations in the immersed tunnel, and an insignificant change occurs in the illuminance of tunnel top and side walls, with the relative error of less than 5%, and the performance degradation of lighting installations in the immersed tunnel has a great impact on the tunnel pavement illuminance.

2.2 Small target recognition

The nature of traffic safety is that drivers can drive correctly according to the tunnel obstacles in front of them [30, 31]. In this paper, the small target recognition test was carried out for lighting installations under different performance degradation conditions to analyze the recognition of small targets by different drivers, which was the basis for traffic safety evaluation. The length of the LED lighting fixture installed in the test tunnel is 100m, and its light distribution curve, layout height, spacing and position are the same as those in the simulation. As recommended by the International Commission on Illumination, a small cube with the size of 0.2 m × 0.2 m × 0.2 m and the surface coefficient of 0.2 was used as the small target [32]. The testing personnel recognized small targets under the following conditions: different distances and various tunnel pavement luminance environments. The visual height of the test personnel is 1.2 m. The components of recognition are shown in Table 2.

TABLE 1 Comparison between measured illuminance results and simulated illuminance results.

Indicator	Average illuminance	Uniformity ratio of illuminance	Longitudinal uniformity
Measured result	208.1	0.52	0.94
Simulation result	226.4	0.53	0.96
Absolute error	18.3	0.01	0.02
Relative error	0.08%	1.93%	2.13%

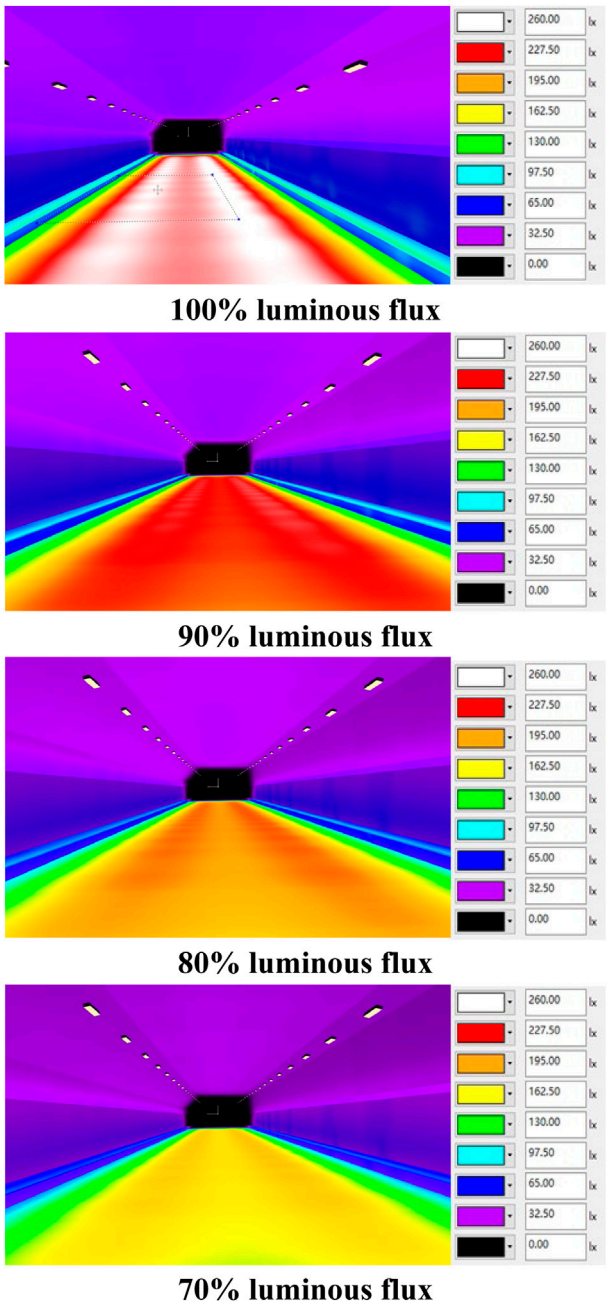


FIGURE 2
Results of luminance simulation of lighting installations in an immersed tunnel.

In consideration of the speed limits in the immersed tunnel, the stopping sight distance at different speed limits, including 100 km/h, 80 km/h and 60 km/h was used as the optimal obstacle avoidance distance, and testing personnel made observations at 158m, 100 m and 56m; the stepless dimming method was used to adjust the pavement luminance in a tunnel and simulate the performance degradation of lighting installations in an immersed tunnel. A total of 29 testing personnel aged between 22–25 were involved, with the visual acuity of above 1.0 and without other vision problems, and the visual height was 1.2 m. After the completion of small target recognition by testing personnel, the small target visibility was identified according to the table of recognition components; the results are shown in Figure 3.

2.3 Traffic safety factor

The results of small target recognition are acquired through the qualitative description of drivers, but the qualitative description is relatively abstract, so it cannot accurately describe the degree of traffic safety and shall be transformed into the quantitative expression of results. Therefore, the results of small target recognition by testing personnel were $y = \{Clear, Relatively\ clear, Fuzzy, Invisible\}$ transformed successively into $\bar{y} = \{4, 3, 2, 1\}$ in this paper. If the recognition result was "clear", the membership degree was 1; if the recognition result was "relatively clear", the membership degree was 0.8; if the recognition result was "fuzzy", the membership degree was 0.01. The large Cauchy distribution and logarithmic functions were used as the membership functions with the method of continuous quantization, to get the expression of traffic safety factor:

$$f(x) = \begin{cases} \left[1 + 1.109\,(x - 0.894)^{-2}\right]^{-1} & 1 \le x \le 3 \\ 0.695 \ln x - 0.036 & 3 \le x \le 4 \end{cases} \quad (1)$$

The results of small target recognition by 29 testing personnel in different conditions were put into the traffic safety expression to get the traffic safety factors in different conditions; the results are shown in Table 3.

3 Traffic safety assessment model of immersed tunnels

At present, the grid method is used to detect the pavement luminance of the lighting installations in immersed tunnels and judge the functional status of the lighting installations in immersed tunnels, but there are problems in such detection, including long interval time, interference with traffic and small scope of detection, so it is difficult to effectively judge them in real time. Therefore, based on the traffic safety factors of immersed tunnels under different degradation conditions, a traffic safety assessment model for immersed tunnels was formed in this paper through the analysis of performance degradation status data of lighting installations in immersed tunnels under the influence of environment and electrical factors.

3.1 Indicator extraction

Given that there are many factors influencing the performance degradation of lighting installations, the multi-sensor technology is used to collect the changes in various index factors of lighting installations, and based on the correlation between performance degradation of lighting installations and various influencing indicators, the index factors with high correlation are acquired and used as the key influencing index factors, and the correlation expression is as follows:

TABLE 2 Component of visual cognition.

Level	Clarity	Detailed description
Level I	Clear	Be able to clearly and directly discover the details, contour and shape of small targets
Level II	Relatively clear	Be able to discover small targets and recognize the contour clearly
Level III	Fuzzy	Not be able to directly discover small targets and recognize the contour fuzzily
Level IV	Invisible	Fail to detect any small target at all

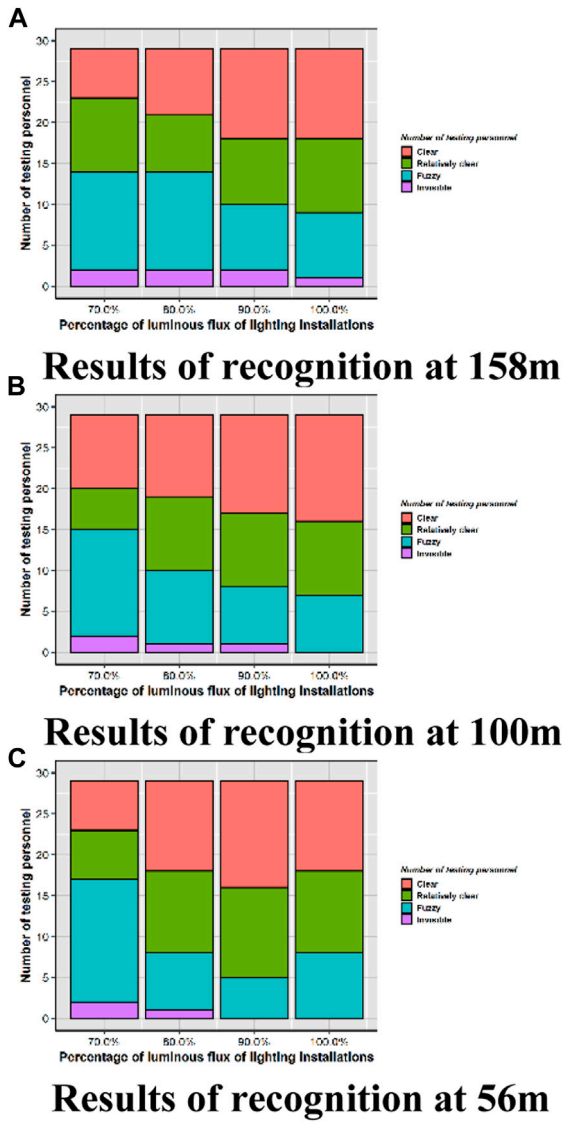


FIGURE 3
Results of small target recognition.

$$r = \frac{\mathrm{cov}(x_a, x_b)}{\sqrt{\sigma_a^2}\sqrt{\sigma_b^2}} = \frac{\sum_{i=1}^{n-1}(x_a - \bar{x}_a)(x_b - \bar{x}_b)}{\sqrt{\sum_{i=1}^{n-1}(x_a - \bar{x}_a)^2 \sum_{i=1}^{n-1}(x_b - \bar{x}_b)^2}} \quad (2)$$

Where r is the correlation coefficient between a and b; σ_a and σ_b are the standard deviations of Indicators a and b; $\bar{x}_a$ and $\bar{x}_b$ are the average values of Indicators a and b.

3.2 Traffic safety assessment model

The correlation among the key indicators influencing the lighting installations in immersed tunnels, the performance degradation of lighting installations and the traffic safety factor is comprehensively analyzed, the dataset of key indicators of lighting installations and traffic safety factors is reconstructed, and the artificial intelligence (AI) algorithm is used to carry out the traffic safety assessment of immersed tunnels. Currently, there are many frequently-used status assessment methods, including grey correlation theory, fuzzy theory, machine learning and neural network [33–35], and with the strong learning ability and generalization ability and the flexible model structure, the neural network is widely used for judgment, prediction and evaluation. The performance degradation process of lighting installations is a time-sequence process, so the long short-term memory (LSTM) neural network in the neural network algorithm was employed in this paper to build a traffic safety assessment model for immersed tunnels [36, 37], and the basic steps are shown.

step1: Build the LSTM neural network, and initialize relevant parameters, such as weight ω and bias (b) of each node in the network, activation function, computational accuracy (l) and number of iterations (n);

step2: Input the dataset (D) in the neural network, abandon the input of information in the forget gate layer, which is unacceptable for the performance degradation indicator data of E&M equipment in the hidden layer at the previous moment, control the input of new indicate data in the input gate layer, determine the data information to be updated, and calculate the output results of the neural network in the output gate layer;

step3: Compare the actual traffic safety factor and the status results of model prediction, and calculate the loss function E;

step4: If the loss function E is less than the setting value, it indicates the completion of training; otherwise, use the gradient descent to update the weight and bias in the neural network, and then return to Step (2) and recalculate it;

step5: After the end of training, obtain the status assessment model for E&M equipment performance degradation.

TABLE 3 Traffic safety factors in different conditions.

Driving speed (km/h)	Lamp degradation (%)	Average illuminance (lx)	Traffic safety factor
100	100	208.1	0.77
100	90	192.4	0.74
100	80	174.3	0.68
80	90	192.4	0.79
80	80	174.3	0.75
60	70	149.7	0.64
...	...	...	...

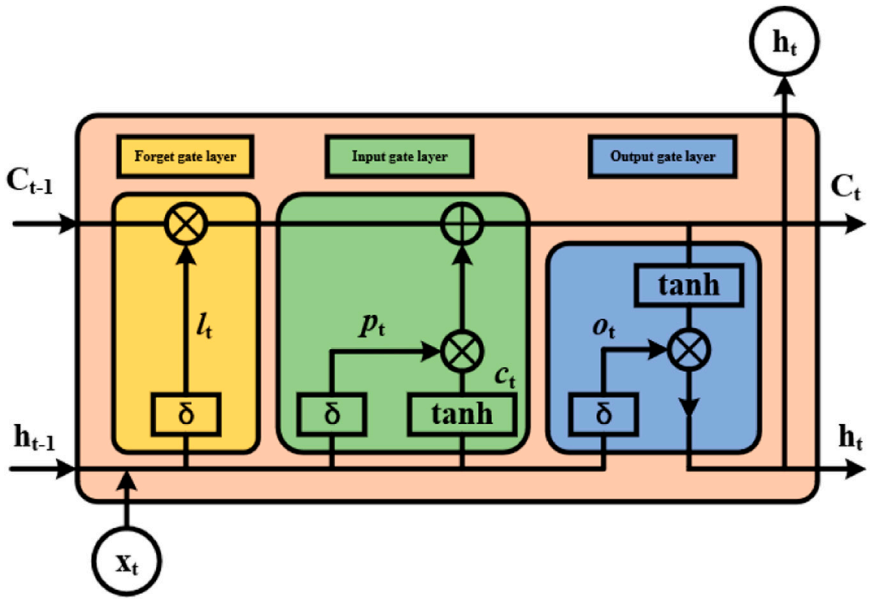


FIGURE 4
LSTM neural network structure.

The LSTM neural network uses the structure containing input gate layer, forget gate layer and output gate layer to substitute for the hidden layer nodes of the traditional neural network, and the network structure is shown in Figure 4.

The forget gate layer reflects the level of acceptance of the neural network to the previous hidden layer status h_{t-1} and the current input status x_t, and its expression is as follows:

$$l_t = \sigma(W_l \bullet [h_{t-1}, x_t] + b_l) \tag{3}$$

Where σ is a sigmoid function; W and b are weight and bias, respectively.

The input gate layer is composed of two parts: Part 1 is to determine the data information (p_t) to be updated (see Eq. 4), and Part 2 is to create the alternative status (c_t) through the tanh layer (see Eq. 5).

$$p_t = \sigma\left(W_p \bullet [h_{t-1}, x_t] + b_p\right) \tag{4}$$

$$c_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c) \tag{5}$$

Based on the output of forget gate layer and input gate layer as well as the network status C_{t-1}, the status expression is updated as follows:

$$\bar{c} = l_t \bullet c_{t-1} + p_t \bullet c_t \tag{6}$$

The output gate layer determines the output value according to the network status, and through the output of tanh layer and sigmoid layer, o_t , as expressed by Eq. 7, determines the output of the hidden layer, h_t , as expressed by Eq. 8.

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t \bullet \tanh(c_t) \tag{8}$$

3.3 Analysis of model results

To realize the evaluation of model results, the root-mean-square error (RMSE) was introduced in this paper to evaluate the model results, and the expression is as follows:

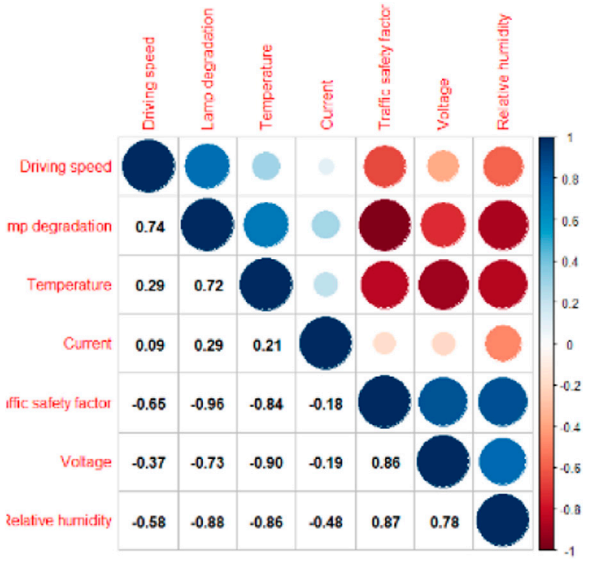


FIGURE 5
Correlation coefficient of each indicator.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

Where y_i and $\hat{y}_i$ are true value and predicted value, respectively.

4 Test verification and discussion

4.1 Data source

Normally, the performance degradation process of lighting installations is characterized by long duration and varying environments. In this paper, the accelerated test data were used to obtain the degradation of lighting installations under different conditions, and combined with the traffic safety factor and the degradation of lighting installations, the data results were reconstructed, with the data size of 1,476 entries.

4.2 Data preprocessing

4.2.1 Data standardization

As one of the frequently-used methods of LSTM to solve the data fitting problems, data standardization is mainly used to eliminate the impact caused by the difference in the order of magnitude between different indicators, with the aim to make weight configuration more reasonable, accelerate data convergence and enhance the accuracy of data analysis, so the z-score method is employed for standardization in this paper, with the following expression:

$$y = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}} \tag{10}$$

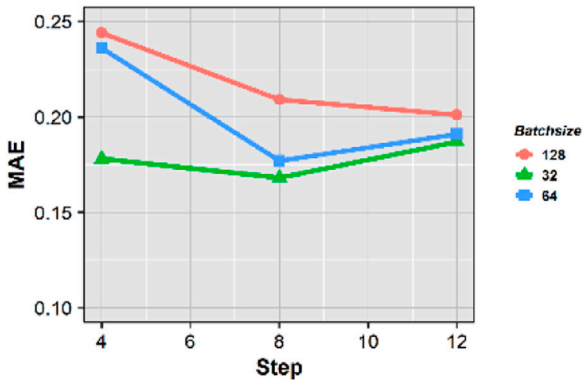


FIGURE 6
MAE value with different steps and batch sizes.

4.2.2 Key indicator extraction

Given that different indicators have different impacts on the traffic safety of immersed tunnels, the analysis of key indicators and the traffic safety assessment of immersed tunnels are carried out to effectively increase the accuracy of assessment results. In this paper, the correlation between various assessment indicators and the traffic safety factor of tunnels was analyzed and the key indicators influencing lighting installations were extracted. The correlation coefficient (r) of various indicators is shown in Figure 5.

It can be seen from Figure 5 that the influencing indicators strongly correlated with the traffic safety factor ($r > 0.7$) include voltage, temperature, humidity and Lamp degradation. Consequently, voltage, temperature, humidity and Lamp degradation were used as the assessment indicators in this paper to realize the traffic safety assessment of immersed tunnels.

4.3 Optimization of model parameters

In the neural network, the mean square error (MSE) or the mean absolute error (MAE) is often used as the loss function, but MSE is sensitive to abnormal values and the process of abnormal values might affect the integrity of actual data, so MAE was used as the loss function in this paper.

When a performance degradation assessment model for lighting installations is built, model optimization shall be carried out. The relevant parameters of LSTM were set, such as sample size (samples), time step (time_steps) and batch size (batch_size), the LSTM assessment models with different times steps and different batch sizes were built, and MAE was utilized for verification; the results are shown in Figure 6.

It can be seen from Figure 6 that when the time step remains unchanged, an increase occurs with the batch sizes, and the MAE value is gradually increased; when the data batch remains unchanged, an increase occurs with the time steps, and the MAE value is increased and then decreased. To acquire a more accurate data model, the time step of 8 and the batch size of 32 were selected as the model parameters in this paper.

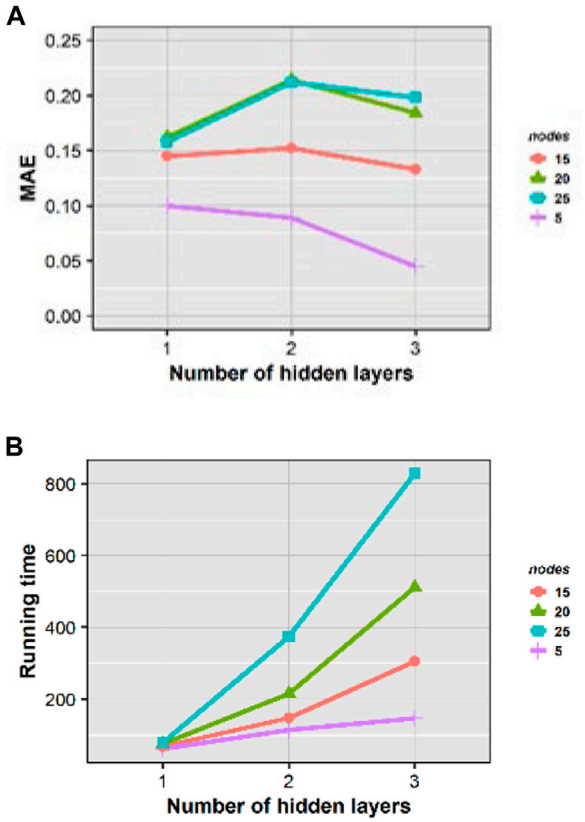


FIGURE 7
MAE values under different hidden layer parameters.

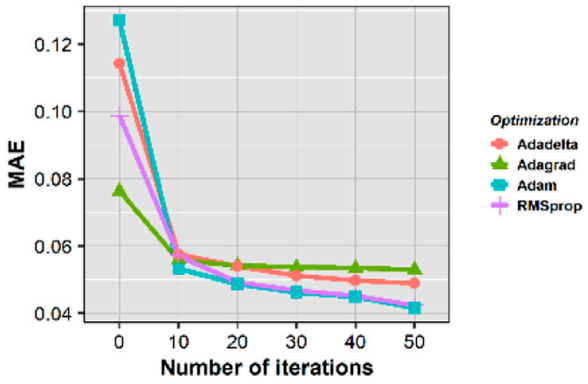


FIGURE 8
MAE values by using different optimizers.

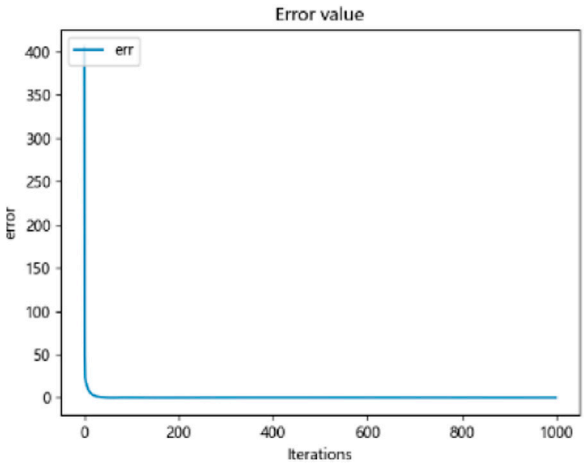


FIGURE 9
Error function value.

The number of hidden layers and the number of nodes are the key parameters of LSTM, so an appropriate selection of the number of hidden layers can improve the accuracy of data results. Model optimization was conducted for the neural network models with different hidden layers (1, 2 and 3) and different hidden layer nodes (5, 15, 20 and 25), and the number of iterations was set as 100. The MAE values under different parameters are shown in Figure 7.

Based on the figure of errors of different LSTM network layers, it can be seen that the minimum assessment error occurs when the number of hidden layers is 3 and the number of hidden layer nodes is 5. If the number of network layers remains unchanged and the number of nodes is increased, the MAE value will be gradually increased; if the number of nodes remains unchanged and the number of hidden layers is increased, the MAE value will be gradually decreased, but an increase in the number of hidden layers will lead to an increase in training and evaluation time. Therefore, in consideration of model accuracy and a reduction in model running time, the number of hidden layers is finalized to be 2 and the number of hidden layer nodes is finalized to be 5 to build the assessment model.

To select an appropriate optimizer, Adagrad, Adadelta, RMSprop and Adam were analyzed and compared in this paper, and the time step was set as 8. The training results are shown in Figure 8.

It can be seen from the diagram that as the number of iterations increases, the loss function of the LSTM model in each optimizer is gradually reduced, and as the model loss function of the optimizer, Adam and RMSprop decline at the fastest pace, and when the number of iterations is 50, the loss function of the model with Adam as the optimizer is 0.0416, which is the minimum value, so Adam is selected as the model optimizer.

4.4 Analysis of model results

In this paper, the dataset of lighting installation tests was randomly divided into a training set and a test set. The training set data accounted for 80% of the dataset, while the test set accounted for 20%. The training set was put into the traffic safety assessment model for immersed tunnels, and its error function values and data prediction results were shown in Figures 9, 10.

It can be seen from Figure 9 that in the first 26 iterations in the process of algorithm iteration, the model error value decreases rapidly, and in the subsequent iteration process, the error value decreases steadily until it is close to zero, and the model fitting process is gradually convergent; in

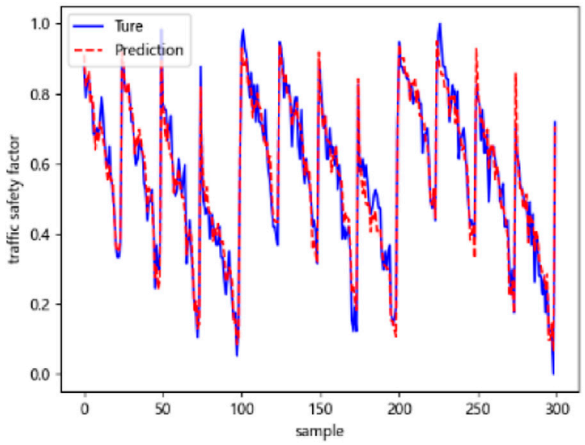


FIGURE 10
Data prediction results.

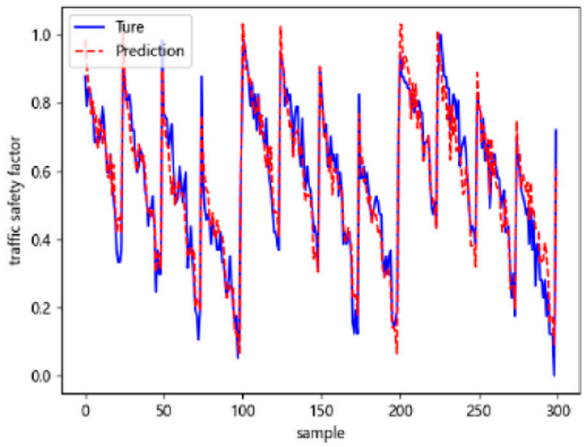


FIGURE 11
RF model calculation results.

Figure 10, the true value and predicted value in the training fitting results show a consistent trend of variations, with a relatively small error value, indicating that the model prediction results are better.

To ensure the accuracy of the traffic safety assessment model for immersed tunnels, the random forest (RF) and recurrent neural network (RNN) models were separately compared with the LSTM model in this paper. Test data are separately put into the RF and RNN models to compare the calculation results of different models (as shown in Figures 11, 12); the results of RMSE, R square and running time are shown in Table 4.

It can be seen from Figures 11, 12 and Table 4 that, compared with the RF and RNN models, the LSTM model can assess the traffic safety of immersed tunnels more effectively, and its RMSE is relatively small and the accuracy is relatively high (up to 95%), indicating that the predicted values of the LSTM model is closer to the actual value in the process of data prediction; the running time of the LSTM model is longer than that of other models, but the running time does not exceed 1min, so the model prediction time can completely meet the demand for the routine traffic safety assessment of tunnels.

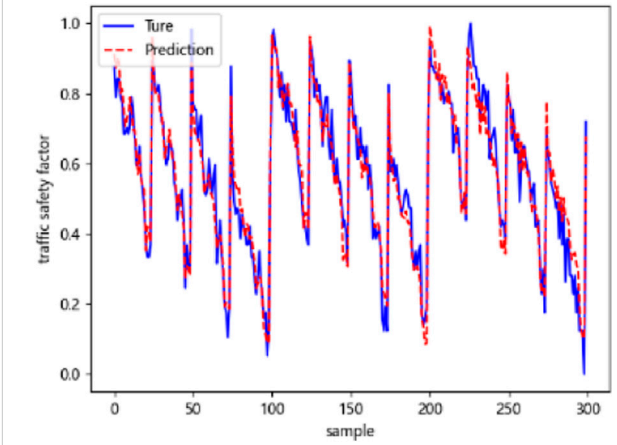


FIGURE 12
RNN model calculation results.

TABLE 4 Comparison of calculation results of different models.

Model	RMSE	R Square	Running time (s)
RF	7.563	0.89	27.42
RNN	2.581	0.91	33.24
LSTM	1.029	0.95	47.81

4.5 Discussion

The quality of luminous effect of lighting installations in immersed tunnels has an important impact on the traffic safety of drivers and passengers. From the perspective of visual needs of drivers and passengers, the traffic safety assessment factor of immersed tunnels was established in this paper by simulating the characteristics of the tunnel lighting environment under different performance degradation conditions of tunnel lighting installations and according to the results of small target recognition by drivers and passengers; combined with the key indicators influencing the performance of lighting installations, the impact of the changes to the operating environment of lighting installations in immersed tunnels on traffic safety was explored in this paper. In the performance degradation simulation of lighting installations, it is difficult to simulate the production quality of different lighting installations, it is believed that the law of performance degradation of the same batch of lighting installations is identical, the parameters of lighting installations set in numerical simulation are identical, and the performance degradation simulation of luminous flux is identical. As a result, there might be an unavoidable error between the actual impact of performance degradation of lighting installations on the pavement lighting environment and the simulated performance degradation results. In a field test, the small target recognition test under dynamic traffic is not carried out for the sake of the safety of drivers and passengers when they dynamically recognize small targets, and testing personnel are stationary relative to a small target and recognize the small target in a physical tunnel, i.e. the surrounding

environment is relatively stationary. There is a difference in the ability of human eyes to dynamically recognize and statically recognize objects, so the results of this study can offer relevant support to traffic safety in real tunnels.

Through comprehensively considering the impact of the operating environment and performance degradation of lighting installations and the traffic safety, a data-driven traffic safety assessment method of lighting installations in immersed tunnels is proposed in this paper, and based on the analysis of the impact of the performance degradation of lighting installations in immersed tunnels on the pavement luminance and the recognition of drives and passengers, the traffic safety factor of immersed tunnels was put forward in this study to evaluate the safety of lighting installations; a real-time kinematic traffic safety assessment model for immersed tunnels was built on the basis of the performance degradation of lighting installations in different environments and under different electrical indicators, and the traffic safety factor of immersed tunnels was dynamically predicted according to the relevant parameters, such as voltage, temperature, humidity and Lamp degradation. In this paper, the traffic safety factor, at which 50% of the testing personnel could clearly recognize small targets, was used as the boundary value, i.e. 0.6304, and if the traffic safety factor was less than 0.6304, it indicated that the lighting performance of the immersed tunnel is poor, and maintenance and repair measures could be taken to maintain the lighting installations, e.g. regular replacement and cleaning; if the traffic safety factor was higher than 0.6304, the luminous flux could be reduced properly to meet the requirements for traffic safety and lower the costs of lighting.

5 Conclusion

In this study, the traffic safety factor was proposed through the analysis of the impact of the performance degradation of lighting installations in immersed tunnels on the recognition status of drivers and passengers, and a real-time kinematic traffic safety assessment model was built according to the key index factors influencing the lighting performance degradation of immersed tunnels, with the aim to ensure the operating safety of immersed tunnels. The research conclusions mainly include:

1) The performance degradation of lighting installations is of great significance to the visual clarity of drivers and passengers and even to the traffic safety through the simulation of the changes in the tunnel pavement luminance and the small target recognition by drivers and passengers under the performance degradation of lighting installations in immersed tunnels. With the performance degradation of lighting installations, the small target recognition of drivers and passengers became weaker and weaker and the traffic safety factor became small, and even it is difficult to detect the existence of small targets.
2) Combined with the key indicators influencing the performance degradation of lighting installations, a real-time kinematic traffic safety assessment method for tunnels was proposed. The traffic safety factor was predicted according to the key indicators influencing the performance degradation, and the R square and RMSE of this model were superior to those of RF model and RNN model, so it could better predict the traffic safety factor of tunnels; meanwhile, the speed was predicted to be within 1min, meeting the requirements for the real-time kinematic assessment of immersed tunnels.
3) In this study, the traffic safety factor could quantitatively evaluate the performance status of lighting installations in immersed tunnels, and according to the small target recognition by testing personnel, the traffic safety boundary condition was put forward, i.e., the traffic safety factor was 0.6304, providing a new direction for the maintenance and repairs of the immersed tunnel management entity.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding authors.

Author contributions

MY wrote the original manuscript and designed algorithm. JC, HD, and SL proposed the idea, supervised the research work, and revised the manuscript. MY and JC discussed and analyzed the results. All authors contributed to the article and approved the submitted version.

Funding

The present work was supported by the National Key R&D Program of China (No.2019YFB1600702).

Conflict of interest

Authors MY, HD, and JC were employed by the company China Merchants Chongqing Communications Technology Research and Design Institute Co, Ltd. Author SL was employed by the company Guangxi Xinhengtong Expressway Co, Ltd.

Publisher's note

References

1. Xiang-sheng C, Zhi-hao X, Xiao-hua B. Challenges and technological breakthroughs in tunnel construction in China[J]. *China J Highw Transport* (2020) 33(12):1–14.

2. Guo H, Yan Y, Ding H (2023). Development and application of automatic monitoring equipment for differential deformation of element joint in immersed tunnel[J]. *Front Phys* 11:94.

3. Cantisani G, Di Mascio P, Moretti L. Comparative life cycle assessment of lighting systems and road pavements in an Italian twin-tube road tunnel. *Sustainability* (2018) 10(11):4165. doi:10.3390/su10114165

4. Peña-García A, Nguyen TPL. A global perspective for sustainable highway tunnel lighting regulations: Greater road safety with a lower environmental impact. *Int J Environ Res Public Health* (2018) 15(12):2658. doi:10.3390/ijerph15122658

5. Miller R. *Shaping the future for immersed tunnels[J]*. (2016).

6. Montero JI, Antón A. Buoyancy driven ventilation in tropical greenhouses[J]. *Acta Horticulturae* (2000)(534) 41–8. doi:10.17660/actahortic.2000.534.3

7. Hirakawa S, Toeda K, Sugawara T, Sakamoto S. Maintenance factor in tunnel lighting installation. *J Light Vis Environ* (2013) 37(1):28–36. doi:10.2150/jlve.ieij130000495

8. Sung K. W. A study of the roust degradation model by analyzing the filament lamp degradation data[J]. *Proc Korean Soc Automotive Eng* (2012) 20(6):132–9.

9. Chiao CH, Wang WY. Reliability improvement of fluorescent lamp using grey forecasting model. *Microelectronics Reliability* (2002) 42(1):127–34. doi:10.1016/s0026-2714(01)00243-8

10. Ministry of Transport of the People's Republic of China. *JTG_H12-2015 technical specifications of maintenance for highway tunnel[S]*. Beijing: People's Communications Press (2015).

11. Cengiz M. The relationship between maintenance factor and lighting level in tunnel lighting. *Light Eng* (2019) 75–84. doi:10.33383/2018-115

12. Lee MW, Park KY, Kim PY, Park YJ, Kim H. A study on the calculation of maintenance factor(MF) of tunnel lighting in expressway considering the actual installation and maintenance conditions. *J Korean Inst Illuminating Electr Installation Eng* (2013) 27(3):7–15. doi:10.5207/jieie.2013.27.3.007

13. Dong W. Design and implementation of maintenance system for tunnel lighting equipment[J]. *Mod Manufacturing Tech Equipment* (2019).

14. Ruifeng L, Aiqiang X, Weichao S. Feature selection algorithm recommendation method for avionics based on meta-learning[J]. *Syst Eng Electron Tech* (2021) 502(07): 2011–20.

15. Qingbing Z, Shiyuan W, Xiaofei Z. Research on early warning method of key equipment in nuclear power plant based on parameter autoregressive algorithm[J]. *Nucl Power Eng* (2021) 249(06):209–14.

16. Xi W. Yang Yongping. Dual-source system of co-generation of internal combustion engine and building load matching and operating characteristics analysis[J]. *Proc Chin Soc Electr Eng* (2014) 481(02):217–24.

17. Du W, Fu Q, Wang HF. Strong dynamic interactions between multi-terminal DC network and AC power systems caused by open-loop modal coupling[J]. IET Generation. *Transm Distribution* (2017) 11(9):2362–74.

18. Heitor R. Impact assessment of virtual synchronous generator on the electromechanical dynamics of type 4 wind turbine generators[J]. IET Generation. *Transm Distribution* (2019) 13(23):5294–304.

19. Omri N, Al Masry Z, Mairot NS, Giampiccolo S, Zerhouni N. Towards an adapted PHM approach: Data quality requirements methodology for fault detection applications. *Comput Industry[J]* (2021) 127:103414–3. doi:10.1016/j.compind.2021.103414

20. Compare M, Bellani L, Zio E. Reliability model of a component equipped with PHM capabilities. *Reliability Eng Syst Saf* (2017) 168:4–11. doi:10.1016/j.ress.2017.05.024

21. Yuliang D, Yaqiong LI, Haibin C. Real-time health condition evaluation on wind turbines based on operational condition recognition[J]. *Proc CSEE* (2013) 33(11):88–95.

22. Li W, Wei H, Qi G. A fast image dehazing algorithm for highway tunnel based on artificial multi-exposure image fusion[C]//IOP Conference Series: Materials Science and Engineering. *IOP Publishing* (2020) 741(1):012038.

23. X Zhen, L Gang, X Daqing Comprehensive analysis of the operating status of the electromechanical equipment of highway tunnels based on the Internet of Things technology[J]. *China Transportation Information Technology*, 2019(02): 142–3.

24. Zhang Z, Chu C, Zhu L. A method of highway electromechanical system facilities maintenance quality assessment based on AHP and fuzzy theory[C]//international conference on transportation information & safety (2011).

25. Yinli J, Lin W, Yang L. Structural division and effectiveness evaluation of highway tunnel electromechanical system[J]. *Mod Tunnelling Tech* (2016) 368(03):47–53.

26. Cui HJ, Zhu CZ, Wang L. Performance assessment model for highway electromechanical system. *Adv Mater Res* (2013) 706-708(1):892–6. doi:10.4028/www.scientific.net/amr.706-708.892

27. Liwei Z. Real-time assessment technology of operational risk of highway tunnel electromechanical system[J]. *Highway* (2017) (12): 176–81.

28. Jianping Z, Yu Z, Wenqing Z. A new life prediction model for optoelectronic products and its application [J]. *J Opt* (2018) 38(2):7.

29. Chinchero HF. Alonso J M, hugo O T. *LED lighting Syst smart buildings: a review [J]* (2020).

30. Falkmer T, Gregersen NP. A Comparison of eye movement behavior of inexperienced and experienced drivers in real traffic environments. *Optom Vis Sci Official Publ Am Acad Optom* (2005) 82(8):732–9. doi:10.1097/01.opx.0000175560.45715.5b

31. He S, Bo L, Pan G, Wang F, Cui L. Influence of dynamic highway tunnel lighting environment on driving safety based on eye movement parameters of the driver. *Tunnelling Underground Space Tech* (2017) 67:52–60. doi:10.1016/j.tust.2017.04.020

32. CIE Technical report. *Guide for the lighting of road tunnels and underpasses[R]* (2004). 88–2004.

33. Yuyan W, Bolin L, Chen P, Jun L, Yumin Y. Research review of recurrent neural networks [J]. *J Jishou Univ (Nat Sci Ed)* (2021) 42(1):41–48.

34. 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

35. 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* (2022) 206:112346. doi:10.1016/j.measurement.2022.112346

36. Qi G, Zhang Y, Wang K, Mazur N, Liu Y, Malaviya D, et al. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. *Remote Sensing* (2022) 14(2):420. doi:10.3390/rs14020420

37. Francisco O, DanielConvolutional RD, Recurrent Neural LSTM. Networks for multimodal wearable activity recognition[J]. *Sensors* (2016) 16(1):115–8.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 23 March 2023
DOI 10.3389/fphy.2023.1165727

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Guanghao Zhang,
Institute of Electrical Engineering (CAS), China
Ruijuan Chen Chen,
Tianjin Polytechnic University, China
Gang Hu,
Buffalo State College, United States

*CORRESPONDENCE
Gui Jin,
tjjingui@126.com
Nan Liu,
natasha0902@sina.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 14 February 2023
ACCEPTED 06 March 2023
PUBLISHED 23 March 2023

CITATION
Xu R, Zhuang W, Bai Z, Wang F, Chen M, Liu N and Jin G (2023), A pilot study on intracerebral hemorrhage imaging based on electrical capacitance tomography. *Front. Phys.* 11:1165727. doi: 10.3389/fphy.2023.1165727

A pilot study on intracerebral hemorrhage imaging based on electrical capacitance tomography

Rui Xu[1†], Wei Zhuang[2†], Zelin Bai[2], Feng Wang[2], Mingsheng Chen[2], Nan Liu[3*] and Gui Jin[2*]

[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: Intracerebral hemorrhage (ICH) is a devastating disease with high rates of mortality and disability. The survival rate and postoperative outcome of ICH can be greatly improved through prompt diagnosis and treatment. CT and MRI are now the gold standards for the diagnosis of ICH, but they are not practical for use in pre-hospital emergencies or at the bedside monitoring.

Methods: Based on the earlier research of ICH detection with a single parallel plate electrode sensor, we developed a 16-electrode Electrical Capacitance Tomography (ECT) system for two-dimensional tomographic imaging of ICH in this study. A 5-layer spherical numerical model and an *ex vivo* porcine physical model of ICH were created for ECT simulation imaging and actual imaging, respectively, to assess the feasibility of this ECT for ICH imaging.

Results: The bleeding circles were easily seen in the image reconstruction in numerical imaging. In *ex vivo* imaging, the existence of bleeding was also more clearly shown with the ECT system; however, the position of the bleeding reconstructed in the image was offset by 3 mm from the real site.

Discussion: The study analyzes the causes of this discrepancy and discusses the steps that may be taken to rectify it. Overall, the simulation and *ex vivo* experimental trials validated the potential of ICH imaging with the ECT method; however, further work is required to increase the performance of the ECT and a more advanced imaging reconstruction algorithm is urgently needed for ICH imaging.

KEYWORDS

intracerebral hemorrhage, ECT, image reconstruction, tomography, bleeding

1 Introduction

Spontaneous intracerebral hemorrhage is caused by the rupture of blood vessels in the brain parenchyma. It is considered the most serious type of acute stroke because of its emergency, dangerous condition, high morbidity, and mortality. ICH accounts for 2.8 million fatalities annually, with the annual incidence rate of 4.1% [1]. The incidence rate of hemorrhagic stroke in China was 126.34 per 100,000 person-years [2], as reported in the 2018 China Stroke Prevention and Treatment Report. The survival rate and postoperative outcome of ICH may be greatly enhanced with early identification and

treatment [3]. Currently, CT and MRI scans are the most important methods for detecting ICH; however, a significant amount of time is lost between transporting the patient to the hospital, performing the CT examination, and receiving the final result, resulting in a missed window of opportunity to treat the condition. The second is that an excessive postoperative bleeding after a hemorrhage often occurs in clinics [4]. Therefore, a portable, low-cost, and fast technology for detecting ICH is urgently needed.

Methods such as electrical impedance tomography (EIT) and magnetic induction tomography (MIT) were developed to detect ICH by measuring electrical resistivity and conductivity with a multi-sensor surrounding a head from all directions to reconstruct the resistivity and conductivity distributions in the brain over the cross section, as blood has different values for these parameters than the rest of the brain tissues [5, 6]. EIT and MIT are promising tomography technologies because of their advantages such as being non-intrusive and non-invasive. Cerebral stroke imaging with EIT and MIT methods has been researched more often and has produced some results, although there are still some issues [7–9]. Due to the relatively high electrical impedance of the skull, there is a considerable attenuation of the excitation current in EIT. Second, EIT requires the connection of electrodes with the scalp, which results in a very large contact impedance. These problems lead to low sensitivity of EIT to brain tissue imaging. As for MIT, the induced magnetic field generated in biological tissues exposed to an excitation field is negligible because of the poor conductivity of biological tissues (0.1 S/m–2 S/m) [10]. Furthermore, the conductivity of blood is not noticeably different from those of other brain tissues [11]. Because of these two factors, MIT has relatively poor sensitivity for visualizing a brain hemorrhage.

Studies of the dielectric properties of brain tissues show that the permittivity of blood is much higher than that of other tissues. At 1 MHz, the permittivity of blood, gray matter, and cerebrospinal fluid is 3,000, 990, and 108, respectively [11]. Though the permittivity of all brain tissues drops with frequency, the permittivity of blood is uniformly larger. Therefore, in theory, imaging the permittivity distribution is preferable to imaging the conductivity distribution of brain tissues for detecting cerebral hemorrhage. We have measured the change in the permittivity in the process of an ICH with a single-channel previously. First, we employed a transmitting coil and a receiving coil, based on the MIT principle, to measure the real part change in $\Delta B/B$ (the induced magnetic field ΔB is relative to the excitation field B) during ICH in a rabbit's head, that is, to measure the change in the brain permittivity, as the information of the permittivity of the measured object is stored in the real part of $\Delta B/B$ [12], as deduced by Griffiths et al [6]. Changes in the real part of $\Delta B/B$ were found to be approximately proportional to the volume of the blood injected [12]. As the real part of $\Delta B/B$ is very small, it is extremely hard to measure and is constrained by a wide variety of factors. Next, we used a parallel-plate capacitor to directly measure the capacitance of the head during hemorrhage, with the resulting changes in capacitance reflecting the corresponding changes in hemorrhage volume [13]. The capacitance of the parallel-plate increased with increasing blood injection volume, as shown in animal experiments [13]. These two experiments suggest that it is indeed possible to reflect the amount of hemorrhage by detecting changes in the brain permittivity. Based on these results, this paper tries to use a multi-parallel-plate electrode to measure the capacitance of the brain in each projection and attempts to realize two-dimensional tomographic imaging of cerebral hemorrhage with the capacitance data and a reconstruction algorithm. This imaging method is known as ECT, which is based on capacitance measurements from a multi-electrode sensor surrounding an object (such as a pipeline or a vessel containing gas, oil, and water in the industry), and has been under development for more than a decade [14]. ECT has been widely used in multiphase flow measurements in the oil industry and fluidized bed measurements in the pharmaceutical industry [15, 16]. Apart from industrial applications, ECT is used to detect breast cancer and brain tumors and to image brain activity [17–19]. W.P. Taruno et al. used a sensor with hemispherical electrode distribution for breast cancer detection based on the fact that the permittivity of the cancerous breast cells is higher than that of healthy breast tissue [17]. They designed an actual phantom in which a paraffin wax ($\varepsilon_r = 1$) imitates human breasts and a rubber ball ($\varepsilon_r = 80$) imitates cancer cells. The phantom was used for three-dimensional ECT imaging. The results showed that the malignant cancerous cells were successfully reconstructed. ECT has also been applied for the detection of brain tumor where abnormal electrical activity around the tumor area is detected [18]. Five patients suffering from ependymoma, oligodendroglioma, craniopharyngioma, germinoma pineal, and cerebellopontine angle tumors were detected using the three-dimensional ECT system [19]. The study showed a positive correlation with MRI and CT results. We measured the change in capacitance during cerebral hemorrhage in animals using a single electrode pair and found that the amount of hemorrhage and the change in capacitance were approximately linearly correlated; this established a foundation for future studies on 2D ECT imaging of cerebral hemorrhage. This paper presents the development of a 16-electrode ECT 2D imaging system. Then, a numerical hemorrhage model and an isolated porcine brain hemorrhage model are developed to test the performance of the designed ECT system through simulation imaging and actual measurement imaging and confirm the feasibility of ECT for ICH imaging.

2 Materials and methods

2.1 Principle of ECT

A typical ECT system comprises three main units: a multi-electrode sensor, sensing electronics, and a computer for hardware control and data processing, including image reconstruction. All electrodes, usually 8 or 12, are mounted around a pipe or vessel, and the capacitance values between all single electrode combinations are measured. The sensing electronics are used to switch one electrode being connected from an excitation signal to a measurement circuit and convert the capacitance into voltage signals, which are digitized for data acquisition. The computer controls the system hardware and implements image reconstruction to show the permittivity distribution. The ECT procedure involves the forward problem and the inverse problem [20]. The forward problem is to calculate or measure the capacitance between all electrode pairs in the sensor. For a complete measurement process, one of the electrodes is selected in turn as the excitation electrode and others as detection electrodes to obtain the capacitance data between all electrode pairs. With this measurement strategy, the number of independent capacitance measurements is

$$M = \frac{K(K-1)}{2} \tag{1}$$

where K is the number of electrodes. For a 16-electrode sensor, 120 independent capacitance measurements can be measured from different electrode pairs. Capacitance data are a response to the presence of permittivity distribution inside the imaging region and is calculated or measured based on the integration of Poisson's equation (20):

$$C = \frac{Q}{V} = -\frac{1}{V}\int\int_{\Gamma} \varepsilon(x,y)\nabla\phi(x,y)d\Gamma \tag{2}$$

where $\varepsilon(x, y)$ is the permittivity distribution in the sensing field, V is the voltage difference between one electrode pair forming the capacitance C, $\phi(x, y)$ is the potential distribution, and Γ is the electrode surface; $\phi(x, y)$ also depends on the permittivity distribution of $\varepsilon(x, y)$. Therefore, the capacitance C between one electrode pair can be considered a function of permittivity distribution ε; as a result, we get the following equation:

$$C = \xi(\varepsilon) \tag{3}$$

Differentiating both sides of Eq. 3, the change in capacitance in response to a change in permittivity is expressed as follows [21]:

$$\Delta C = \frac{d\xi}{d\varepsilon}(\Delta\varepsilon) + O(\Delta\varepsilon)^2 \tag{4}$$

As the change in permittivity is supposed to be small, Eq. 4 is often simplified to be a linear system. This relationship can be represented by

$$\Delta C = s\Delta\varepsilon \tag{5}$$

where S is the sensitivity matrix. Eq 5 has to be discretized to calculate S and visualize the permittivity distribution. The sensing area is divided into N elements or pixels. The discrete form of (5) can now be expressed as [20]

$$\frac{\lambda}{M\times 1} = \frac{S}{M\times N}\cdot\frac{g}{N\times 1} \tag{6}$$

where λ is the capacitance vector, g is the permittivity vector, that is, the gray level of pixels, 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 finite element simulation. The inverse problem of ECT is to deduce the permittivity distribution from the measured capacitance vector C (Eq. 2). In the discrete form, it is to calculate the unknown g from the known λ using Eq. 6, while S is treated as a constant precalculated matrix [20]. The solving of the inverse problem is also the task of image reconstruction. Because the number of pixels N is usually much larger than the number of capacitance measurements M, Eq. 6 is underdetermined, and the solution is not unique, so the reconstruction algorithms are required to try to find the approximate solution to Eq. 6. In general, the reconstruction algorithms for ECT can be categorized into two groups: non-iterative algorithms and iterative algorithms. Linear back-projection (LBP) is a typical non-iterative algorithm. Common iterative algorithms include iterative Tikhonov regularization, Landweber iteration, conjugate gradient method, and Newton–Raphson method [22].

2.2 ECT sensor and measurement system

Our 16-electrode ECT system (Figure 1A) comprises five main units: an ECT sensor (1), control electronics (2) and (3), an impedance analyzer (4), and a computer (5). As shown in Figure 1B, the ECT sensor consists of 16 square electrodes which are uniformly spaced on a circular base with a diameter of 60 mm. One single electrode shown in Figure 1C is made of a square thin copper film (50 mm * 12 mm) printed on a PCB which includes a solder pad in the middle for soldering electrode leads. The imaging area (Figure 1D) is a circle with a diameter of 60 mm, centered on the electrode circle and evenly subdivided into 812-pixel points. The ECT sensor can provide 16*15/2 = 120 independent capacitance measurements for the inverse calculation of the permittivity for the 812-pixel points. We employed an impedance analyzer (4294A, Agilent Technologies) to measure the capacitance of the electrode pairs. When compared to a conventional capacitance measurement circuit, an impedance analyzer offers many advantages for capacitance measurement, including high measurement accuracy, a shorter development cycle, and the ability to vary the measurement frequency at will (thanks to the analyzer's broad measurement range, which spans 40 Hz–110 MHz). The main drawback of this method is that it takes a long time to measure one capacitance when high precision is needed. One single measurement time of roughly 0.15 s may be achieved when the 4294A impedance analyzer is set to its greatest measurement precision setting (fifth gear). We consecutively measured the capacitance of each electrode pair for three times and then calculated the average values, so the measurement time of all the electrode pairs was 0.15*3*120 = 54 s. By adding the delay time of electrode switching, the time of one measurement period was about 1 min. This imaging speed is not applicable to industrial multi-phase flow measurement but is acceptable for brain hemorrhage imaging, since the hemorrhage in the brain is not expected to quickly alter. The computer (5 in Figure 1A) sends orders to the impedance analyzer through the USB-GPIB interface to trigger capacitance measurements and data collection. The control electronics consist of two parts: relay circuitry and electrode control circuitry (2 and 3 in Figure 1A). Figure 2 depicts the diagram of the electrical connection of the main hardware circuit. The three dashed boxes, 1, 2, and 3 shown in Figure 2, indicate the electrical connections of the three parts, 1, 2, and 3 shown in Figure 1A, respectively. Sixteen electrodes are represented by the dashed boxes from E1 to E16 in 1. Each electrode is wired to the input terminal of two relays connected in series, as shown in the relay connection diagram in Figure 2. As an example, E1 may be linked to any of the three nodes (H, L, or GND) through any of the combinations of logic levels on k1-1 and k1-2 (control terminals of the relay). Omron's electromagnetic G5V-1 relays are employed in this setup. The H and L nodes are connected to the excitation signal output and measurement signal input in the impedance analyzer, respectively. Choosing any one electrode as the excitation electrode or as the measurement electrode and putting the other 14 electrodes to the ground is sufficient to manipulate the control signals of all relays. The H and L nodes are wired to the two ends of the two-port fixture (16047E, Agilent Technologies), as illustrated in Figure 1A. The ground of the control electronics is also wired to the ground of the impedance analyzer. The connection of electrode-controlled circuitry, shown by the dashed box 3 tbox3 in

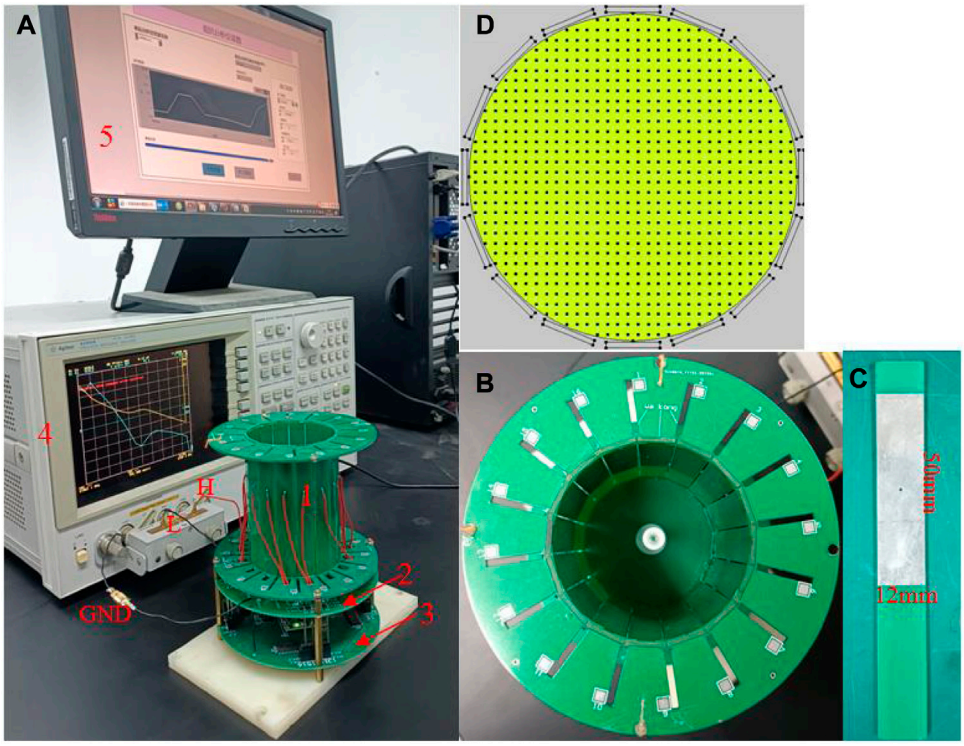


FIGURE 1
Photographs of the ECT measurement system and sensor. **(A)** Measurement system, **(B)** ECT sensor, **(C)** electrode, and **(D)** pixel points of the imaging area.

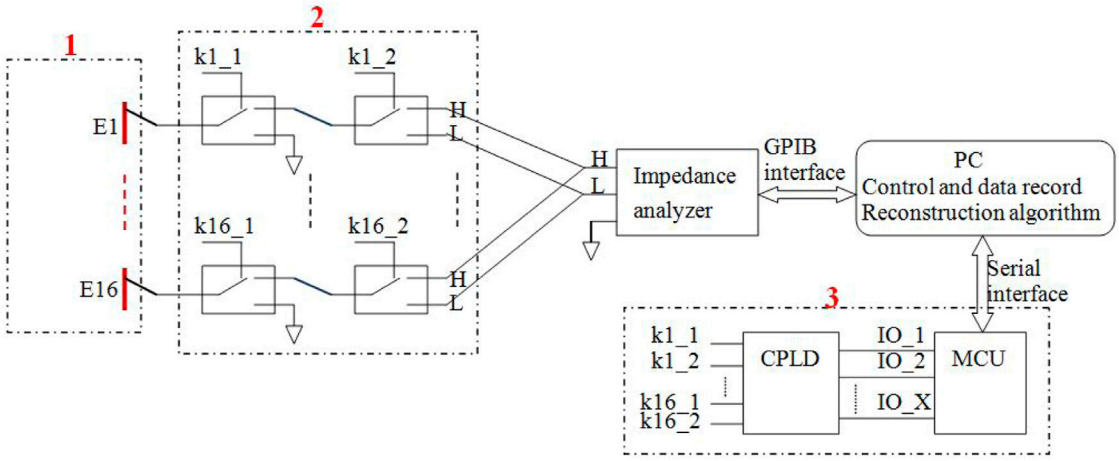


FIGURE 2
Schematic diagram of the electrical connection of the measuring system. The dashed box 1: 16 electrodes. The dashed box 2: The electrical connection of relay circuitry. The dashed box 3: The electrical connection of control circuitry.

Figure 2, mainly consists of a microcontroller (STM32F103C8T6) and a programmable logic chip (CPLD). A decoding circuit in CPLD is used to decode the control signals from MCU and produce the corresponding control signals of all relays to select the excitation electrode and the measurement electrode. The MCU communicates with the PC *via* a serial interface, receives excitation and measurement electrode numbers from the PC periodically, and outputs corresponding control signals to CPLD, which in turn decodes and produces control signals of all relays to connect the corresponding electrodes to H and L nodes. A program for data collection and control was designed based on the LabVIEW platform in PC. The program first sends the excitation electrode and measurement electrode numbers to the MCU, then triggers the impedance analyzer to measure capacitance of the selected electrode

pair, and finally reads the capacitance data from the impedance analyzer. When a measurement cycle is completed, the program creates a data file in the Excel format to save the capacitance data of all electrode pairs. An algorithm in MATLAB calls the measurement data and the sensitivity matrix data to reconstruct an image.

The use of electromagnetic relays for electrode switching instead of analog switches is the primary difference between this ECT measuring system and the standard ECT system. The reason for this is that it has been found experimentally that the off-capacitance of the analog switches is very large (several pF–tens of pF) [23], and for a 16-electrode ECT, at least 16 analog switches are required, and the total analog switching circuit forms a stray capacitance of hundreds of pF. When disconnected from the circuit, one opposite electrode pair of the ECT sensor has a capacitance of approximately 0.3 pF; however, when connected to the analog switching circuit, this value would boost to hundreds of pF. The high stray capacitance greatly reduces the measuring accuracy and resolution of the capacitance induced by the object under test. When an electromagnetic relay is in an off condition, the terminals are disconnected physically, leading to a tiny off-capacitance. The capacitance of 120 electrode pairs in this ECT system was measured to be between 8.936 pF and 11.328 pF when the imaging area is empty. When the impedance analyzer was set to its greatest accuracy, the ECT system could achieve a precision of 0.0006 pF. The frequency sweep measurement test reveals that the ECT system has minimal measurement noise between 1 MHz and 10 MHz. The lower the frequency, the greater the noise, and the higher the frequency, the smaller the noise, but if the frequency is too high, the more the ECT system is affected by external interference, considering that the measurement frequency is chosen to be 3 MHz.

2.3 Image reconstruction algorithm

In this paper, we used the traditional conjugate gradient (CG) method for solving the linear inverse problem. The CG method has a fast convergence rate but is only applicable to a linear system of equations with a symmetric positive definite (SPD) coefficient matrix [24]. However, the sensitivity matrix S for the ECT problem is always non-symmetric and ill-conditioned. To obtain a stable solution, it is first necessary to regularize the sensitivity matrix S. Considering $g(x, y)$ and $S' = (S^T S + \mu I)$, then Eq. 6 can be expressed as [25]

$$S' \cdot g = \lambda' \tag{7}$$

The solution is then solved according to the idea of the CG method.

3 Experimental arrangements

3.1 Simulation arrangement

To check the imaging performance of the ECT sensor and to calculate the sensitivity matrix S (provided for the actual imaging later), simulations were first performed. The simulation was carried out using COMSOL Multiphysics and MATLAB on a PC equipped with an Intel Core i7 processor of 3.40 GHz. First, the same sensor model is created in COMSOL according to the size of the actual ECT sensor in Figure 1. The imaging area is set as a circle with a diameter of 60 mm (as shown in Figure 1D). For the inverse problem, the imaging area is divided into 32×32 grids, with the outer part of the circle removed, which results in 812 pixels inside the imaging area. The sensitivity of each of the 812 pixels is then calculated for each electrode pair. 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. 8, 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 electrode and (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 electrode. Each electrode is set in turn as the excitation electrode, and E_x and E_y are calculated for all 812 pixel points. Finally, the sensitivity matrix for all electrode pairs is calculated according to Eq. 8.

$$S_{ij}(x, y) = -E_{xi} \times E_{xj} + E_{yi} \times E_{yj} \tag{8}$$

This air domain sensitivity matrix is used for both simulation imaging and later actual imaging. Three simple permittivity distribution models (Figures 3A–C) and two ICH models (I1–I2 in Figure 3) are used for numerical simulation. As for simple models, the background material is air, and the relative permittivity of test objects is 4. The diameter of all the small circles in all simple models is 6 mm. The ICH model is based on the actual brain structure but is simplified by dividing it into five layers from the outside to the inside, simulating skull (rose red), cerebrospinal fluid (brown), gray matter (green), white matter (dark blue), and blood (wine red). The relative permittivity of each part in the ICH model is calibrated to the measured values of human brain tissue from the literature (Table 1). The small red circle in the ICH model is used to represent hemorrhage. The coordinate of the center of the imaging area in I1 or I2 is (0 mm, 0 mm). The small red circle in I1 and I2 represents hemorrhaging in the left hemisphere and right hemisphere, respectively. The diameters of the two small red circles are all 10 mm, and the central coordinates are (−9 mm and 7 mm) and (9 mm and 7 mm), respectively. To image brain hemorrhage, we first subtract the calculated capacitance data of I1 or I2 from the reference data, which are the calculated data with no bleeding present (with the red circles removed and the remainder retained).

To assess the quality of image reconstruction, the relative image error and the correlation coefficient between the true model and reconstructed images are used as criteria. The definition of the relative image error and correlation coefficient is shown in Eqs 9, 10, respectively [21]. The lower the image error and the higher the correlation coefficient mean, the better the image reconstruction results.

$$Image\ error = \frac{\|\hat{g} - g\|}{\|g\|} \times 100\% \tag{9}$$

where

$$Correlation\ coefficient = \frac{\sum_{i=1}^{P}(g_i - \bar{g})(\hat{g}_i - \bar{\hat{g}})}{\sqrt{\sum_{i=1}^{P}(g_i - \bar{g})^2 \sum_{i=1}^{P}(\hat{g}_i - \bar{\hat{g}})^2}} \tag{10}$$

is the normalized pixel value reconstructed and g is the normalized permittivity vector of a true distribution in the model. $\hat{g}$ and $\bar{\hat{g}}$ are the mean values.

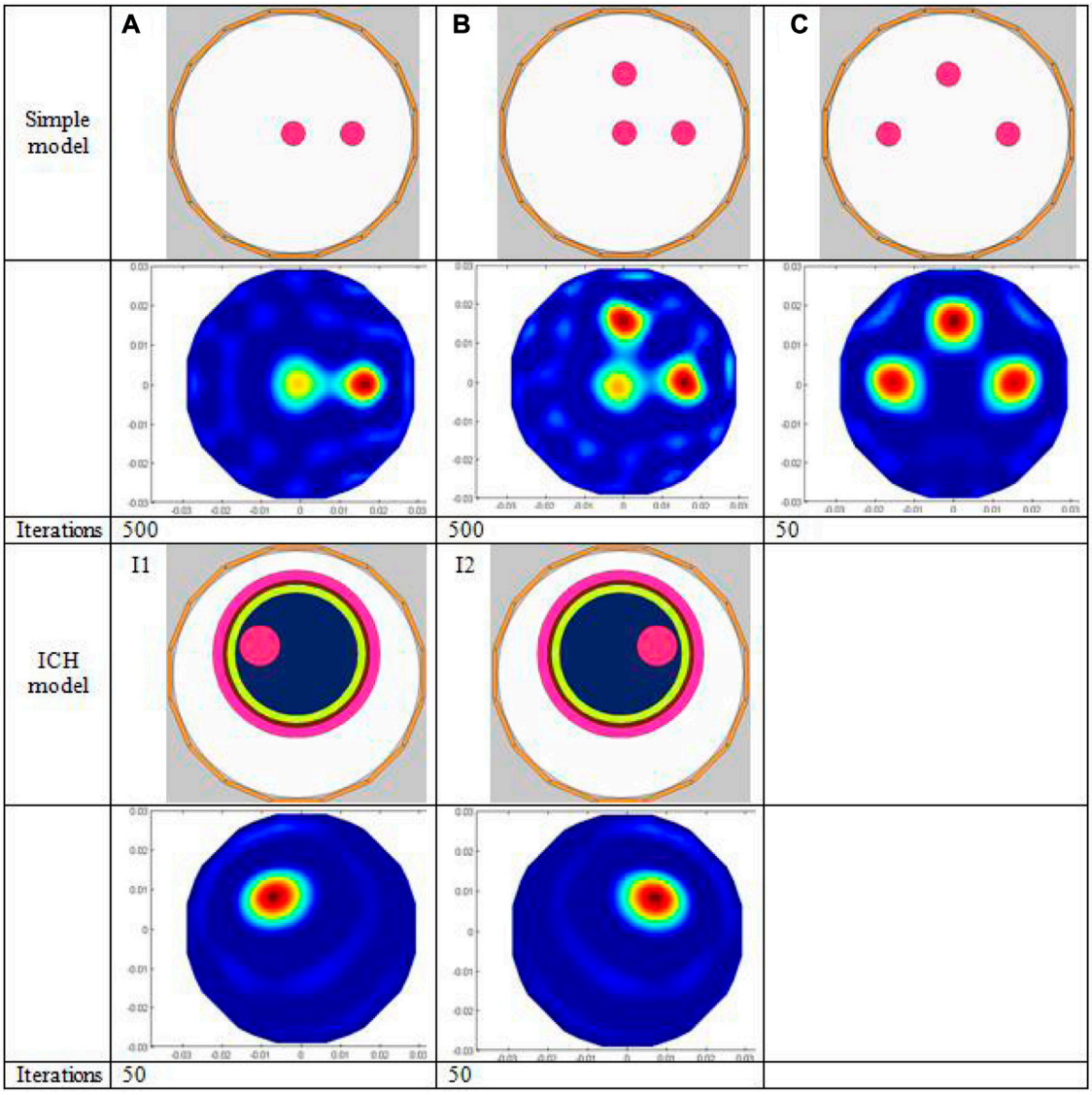


FIGURE 3
Simulation model and imaging results. The first three rows show the simple simulation models **(A–C)** followed by the corresponding imaging results and the number of iterations; the last three rows show the two ICH models (I1, I2), the corresponding imaging results, and the number of iterations, respectively

TABLE 1 Relative permittivity of each part in the ICH model (frequency 1 MHz) [11].

Color	Rose red	Brown	Green	Dark blue	Wine red
Typical tissues	Skull	Cerebrospinal fluid (CSF)	Gray matter	White matter	Blood
Relative permittivity	150	108	991	700	3000

3.2 Actual physical model imaging experiments

As can be seen in Figure 4, the physical model imaging is divided into three parts:Ⅰ, Ⅱ, and Ⅲ. Three models of A1, A2, and A3 in Part Ⅰ are used to image blood at different locations. In models A1 and A2, a plastic tube containing anticoagulant fresh sheep blood (inner diameter 5 mm) is put in the center and 1/2 radius of the imaging region. In A3, two identical blood-filled plastic tubes are positioned at a 1/2 radius apart horizontally. B1, B2, and B3 are three models in Part Ⅱ, in which different solutions are placed in a big cylinder (56 mm inner diameter) and a miniature tube (with a diameter of 10 mm) which is fixed to the position at 1/2 radius of the big cylinder. For B1, water is placed in the large cylinder, while vegetable

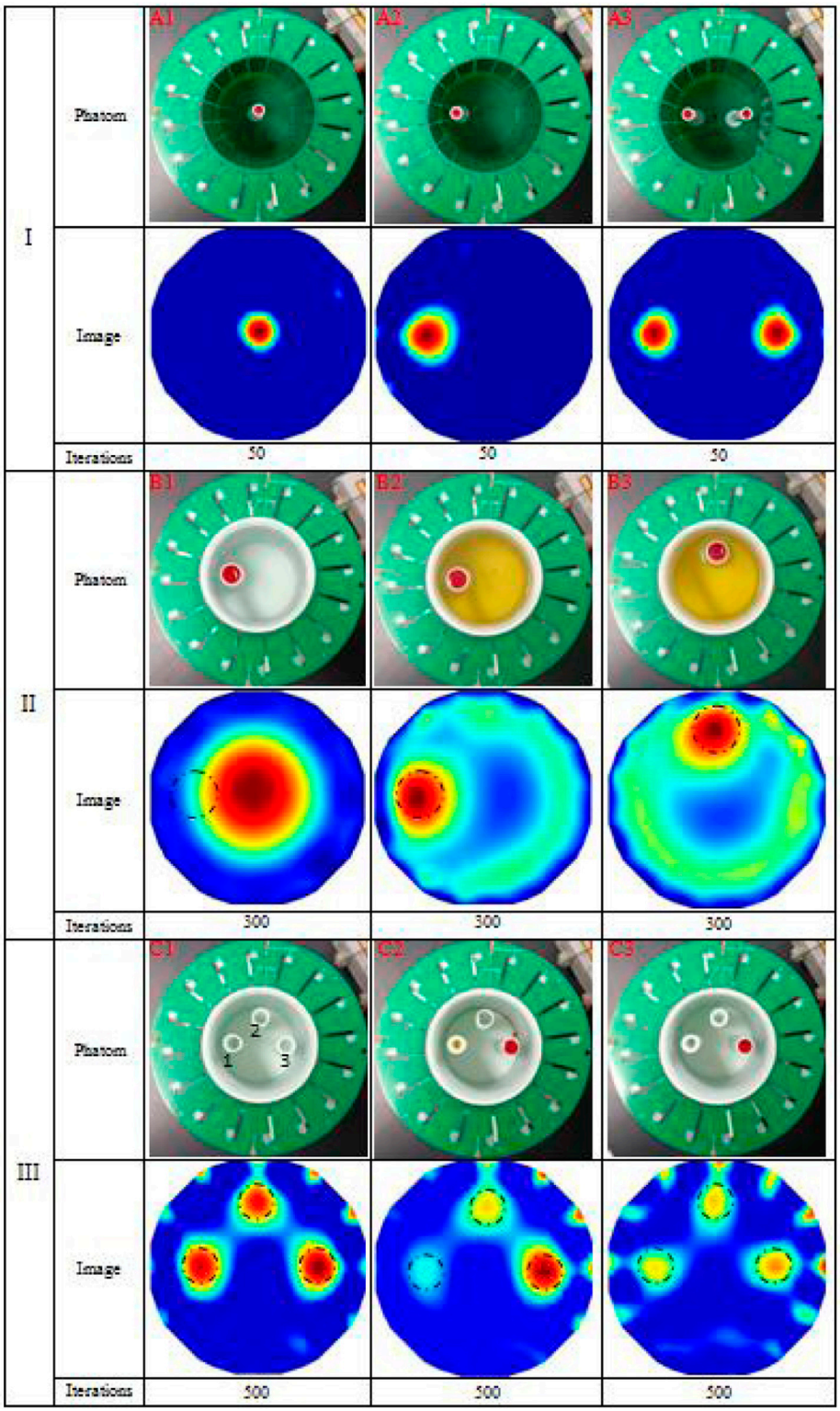


FIGURE 4
Physical experimental models and imaging results.

oil is placed in B2 and B3, and blood is placed inside the small tube for all the three models. First, the big cylinders and small tubes are measured once when empty; the resultant capacitance is used as a reference; next, the cylinders and the small tubes are filled with the appropriate solutions and measured once again; the resulting capacitance is subtracted from the reference capacitance to create an image. Thus, for the B1–B3 model imaging, blood is wrapped in water or vegetable oil. This kind of imaging is a static imaging used

TABLE 2 Relative permittivity of the four solutions (frequency 3 MHz).

Solution	Blood	Water	Alcohol (75%)	Vegetable oil
Relative permittivity	700 (without anticoagulant)	80	40	2–4

mostly to test ECT's capacity to image targets against complicated backgrounds. Models in Part III are C1 –C3. In these models, a tiny tube (inner diameter 8 mm, designated by 1, 2, and 3 in sequence) is positioned within a larger cylinder (inner diameter 56 mm) at 1/2 radius horizontally and vertically. The big cylinder's is empty inside, devoid of any kind of solution. For C1, all three small tubes are filled with water; for C2, three small tubes 1, 2, and 3 are filled with vegetable oil, alcohol, and blood, respectively; for C3, three small tubes 1, 2, and 3 are filled with alcohol, water, and blood, respectively. With three small tubes filled with air, the subsequent measurement data are utilized as reference data. Part III is designed to test the ECT's capacity to image targets with different relative permittivity. The relative permittivity for all of the solutions employed in physical experiments is listed in Table 2, with blood showing the highest value, followed by water, 75% alcohol, and vegetable oil. Note that the blood used in this work has been diluted with the anticoagulant sodium heparin and that early studies showed that the permittivity of this kind of blood is substantially less than 700 listed in Table 2 but greater than that of water.

For imaging of Part II and III, to quantitatively compare the pixel value levels of the reconstructed image of objects under test, an average pixel value parameter AVP is defined as follows:

$$AVP = \frac{\sum_{i=1}^{P} \hat{g}_i}{P} \tag{11}$$

For each reconstructed model image, the circular contour of each part image of the object under test was first manually circled, and then the AVP within that contour was calculated as shown in Eq. 11, with $\hat{g}$ representing the value of each pixel within the contour and P representing the number of pixels within the contour. For B1–B3, the circular contour where the small tube is located is manually circled on each reconstructed image (the diameter being the same for B1–B3), and the AVPs of both the small tube and the remaining part (within the large cylinder) are calculated separately, expressed as AVP_C and AVP_R, respectively, and the ratio of AVP_C: AVP_R is calculated. For C1–C3, the circular contour (with the same diameter), where the three small tubes are located on each reconstructed image, is circled, the AVP of the three small tubes 1, 2, and 3, denoted by AVP_1, AVP_2, and AVP_3, respectively, is calculated, and the ratio of AVP_1: AVP_2: AVP_3 is calculated.

3.3 Isolated porcine brain hemorrhage imaging experiment

To test the feasibility of ECT for imaging actual ICH, we performed *ex vivo* imaging experiments before conducting animal experiments. The market-bought pig brain *in vitro*, a 3D-printed big cylinder (56 mm inner diameter) (with a tiny tube of 11 mm inner diameter printed at its inner 1/2 radius), and a syringe with 10 mL of sheep blood diluted with sodium heparin were prepared, as depicted in Figure 5. The pig brain slices were carefully placed into the larger cylinder, equally piled around the small tube, and firmly yet gently squeezed until their height matched that of the electrode. First, the big cylinder containing the pig brain slices was placed in the ECT sensor co-axially and measured once when the small tube is empty, and the resultant capacitance was utilized as the reference data. Then, the blood in the syringe was slowly injected into the small tube. The new capacitance measurement was taken after the small tube was filled. The imaging data are the difference between the reference capacitance and the data with blood injected. The big cylinder was rotated such that the small tube is located at the left of and below the imaging area, and the imaging test for the aforementioned process was carried out at the two positions (as shown in Figure 5), respectively.

4 Results and discussion

4.1 ECT measurement noise

The measured capacitance data for the 120 individual electrode pairs with an empty sensor field are shown in Figure 6, with a capacitance distribution ranging from 8.936 to 11.328 pF. This empty field capacitance is very small compared to that of the other ECT system and is entirely due to the use of electromagnetic relays. Figure 7 shows the measured capacitance of an adjacent electrode pair over 2 min, with 2 min being exactly the imaging time of one frame. The standard deviation was calculated to be 0.000127 pF.

4.2 Simulation imaging results

Figure 3 displays the simulated imaging results. For simple models A and B, even though all small red circles in the model are set to the same relative permittivity, the pixel value level of the small red circle in the center is substantially lower than that of the small circles positioned at 1/2 radius; thus, the image of the small circle in the center looks a bit hazy. The reason is that the center of the imaging zone for ECT has the lowest sensitivity due to the soft field properties of ECT. Although many advanced reconstruction algorithms in other papers have been found to mitigate this effect successfully, the conjugate gradient approach utilized in this study does not appear to be one of them. The imaging result of simple model C outperforms A and B. Since the radial distance from the center of each of the three smaller circles to the center of the imaging zone is the same (1/2 of the larger circle's radius), the pixel values for each of the smaller circles are roughly the same. The number of iterations in the reconstruction of models A, B, and C are 500, 500, and 50, respectively. For the images of models A and B, the white spot of noise is considerably more pronounced than in the image of model C since the sensitivity of the center is the lowest, and a very high number of iterations are required to show the small circle in the middle. With model C, a decent reconstructed image can be obtained only after 50 iterations. Table 3 displays the image error (%) and the correlation

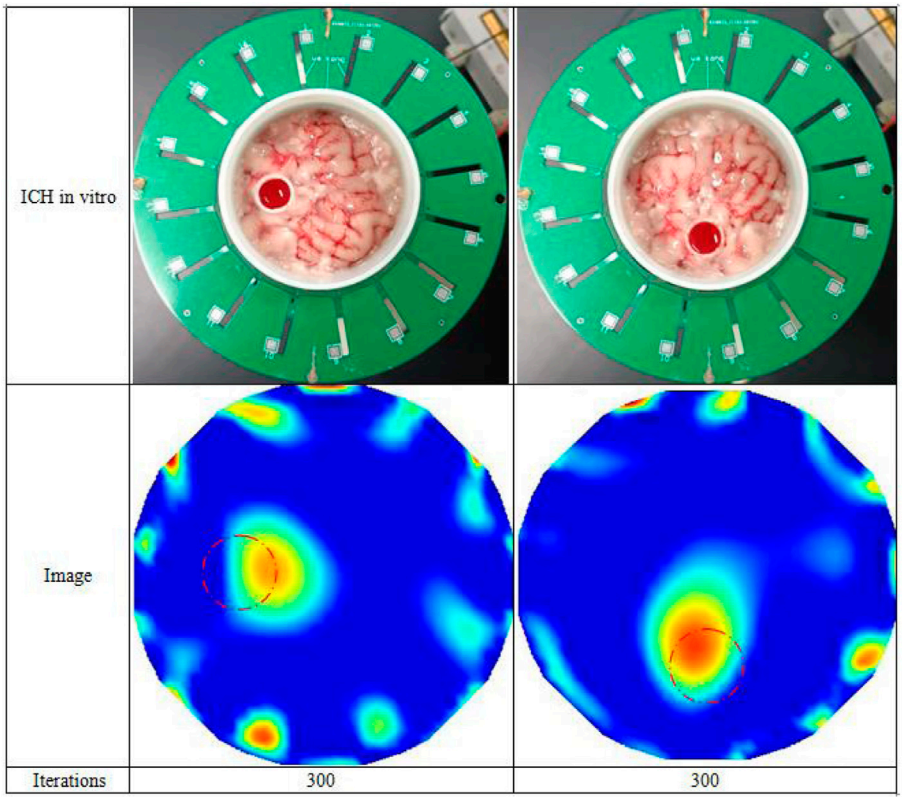


FIGURE 5
Isolated porcine brain hemorrhage models and imaging results.

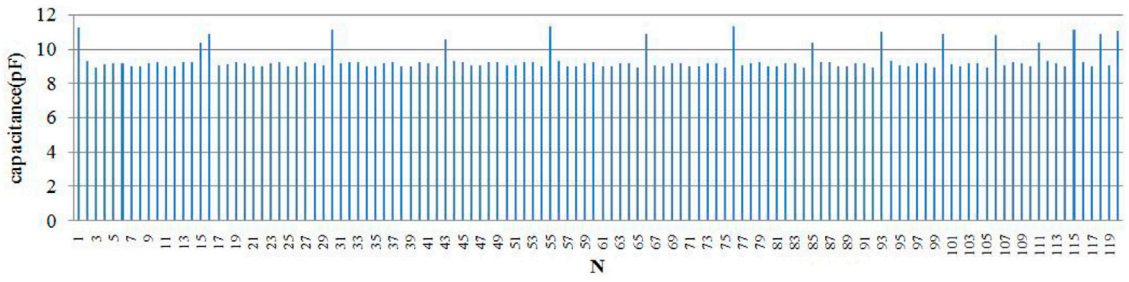


FIGURE 6
Capacitance measurements of 120 independent electrode pairs under an empty field.

coefficient of all models' imaging; the image error (%) of model AB is higher than that of model C, and the correlation coefficient is lower for model AB. The imaging of ICH models I1 and I2 is clear, and the location and size of the hemorrhage are more accurately reflected due to the use of time-difference imaging in large part, in which data calculated with the hemorrhage circle existed is subtracted from data calculated with the hemorrhage circle removed. Second, the diameter of the small red circle in the ICH model is larger than that in the simple model. On the other hand, it shows that ECT can image targets behind multiple layers of material.

4.3 Physical experimental imaging results

Figure 4 displays the imaging outcomes of the physical experiments. Models A1–A3 in Part I exhibit excellent results. The location of the blood in the image is spotted with its location in the model. Nonetheless, the unequal distribution of sensitivity in the imaging area contributes to a much smaller diameter of the blood image in A1 compared to A2 and A3. The imaging of the same object placed in the center has the lowest sensitivity; thus, its image appears smaller than that of the same object placed in other positions. In most cases, ECT can image blood.

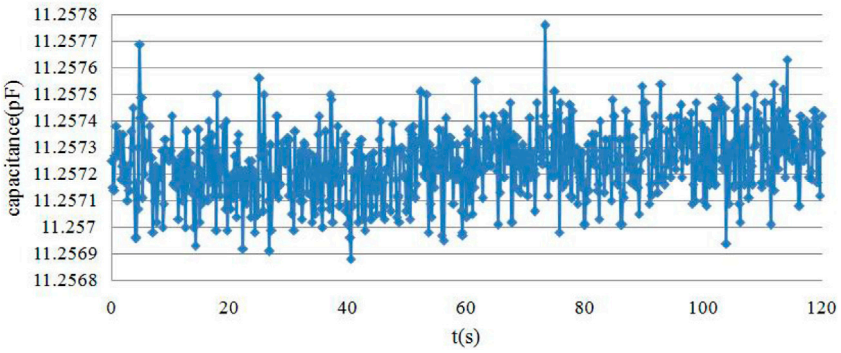


FIGURE 7
Measurement data of one adjacent electrode pair within 2 min.

TABLE 3 Image error (%) and correlation coefficient for simulation results.

	A	B	C	I1	I2
Image error (%)	21.34	25.65	5.62	9.38	9.41
Correlation coefficient	0.8863	0.8244	0.9628	0.9217	0.9135

TABLE 4 Average pixel value ratio of blood to the background for imaging results in Part II.

Phantom	B1	B2	B3
AVP_C: AVP_R		53.27:1	50.59:1

The imaging results of models B1–B3 in Part II is significantly poor. The results of B2 and B3 can still reflect the position and size of blood, but the backdrop appears as huge plaques, mainly because blood is surrounded by other high permittivity backgrounds (not air); thus, it belongs to static imaging rather than differential imaging. The B1 imaging results do not show the presence of the blood portion at all. This is mainly because blood is surrounded by water, and the permittivity of this blood diluted with sodium heparin is not much larger than that of water, as already described. The diameter of tubular blood is only 10 mm, which is only 1/6 of the diameter of water, so the change in capacitance caused by the blood part is very small and submerged in the capacitance change caused by water. Thus, it is difficult to show the presence of blood in the reconstructed image. Images of B2 and B3 show the presence of the blood component, and their location and size are very close to the real model. In addition, part of the reason is that the background of blood wrapped in B2 and B3 is replaced by vegetable oil, and the relative permittivity of vegetable oil is less than 4 (Table 2), which is much smaller than that of blood. For this reason, the imaging can emphasize the existence of the blood portion even though the diameter of tubular blood is much smaller than that of vegetable oil due to the substantial difference in relative permittivity. However, for imaging of models B2 and B3, a large area of plaque appeared in the background outside of the blood image, with areas of low pixel values appearing in the middle near the blood and areas of higher pixel values appearing near the electrode. On the one hand, this is still caused by uneven sensitivity of the imaging area, and on the other hand, it is because the soft field characteristic of ECT will cause the sensitivity distribution of the imaging area to change with the distribution of materials with multiple different permittivities. Table 4 gives the ratio of the average pixel values of the blood area and the remaining background in B2 and B3 images, which is expressed as AVP_C: AVP_R. The ratio of B2 and B3 is closer because in B2 and B3 only the position of blood is different and the rest of the conditions are identical. The average value of B2 and B3 is 51.93:1, which also indirectly reflects that the permittivity of blood is much higher than that of vegetable oil; however, it will not be equal to the actual permittivity ratio of blood to vegetable oil because the imaging will be affected by the volume and position of the objects besides permittivity. The average pixel value ratio in the image of B1 was not determined since blood could not be imaged. This experiment demonstrates that it is exceedingly difficult to conduct static imaging on an object with complex backgrounds with minor changes in permittivity, a recurring bottleneck in the field of electrical tomography.

The imaging findings of C1–C3 in Part III reveal the presence of three tubes of solution more clearly, although there are numerous high-pixel-value patches near the electrodes, which is primarily caused by the noise being increased as a result of the high number of iterations in image reconstruction. For C1, the three tubes are all filled with water and positioned at the same distance from the center, so the images of the three tubes with solution should theoretically be identical. The image of C1 shows that the three tubes' pixel values are roughly comparable. The average pixel value ratio of the three tubes is given in Table 5 as AVP_1: AVP_2: AVP_3 = 1:0.91:1.24, which are close to each other, and the subtle differences may be caused by differences in the placement of the large cylinders. Due to the manual placement of the big cylinder, the center axis of the large cylinder does not exactly coincide with the center axis of the imaging area; therefore, there is a distance discrepancy between the centers of the three tubes and the center of the imaging area. The imaging of C2 reveals three circular regions ranging in color intensity from light to dark from left to right. The solutions in 1, 2, and 3 tubes in C2 are vegetable oil, alcohol, and blood in the order of increasing permittivity, which correlates to the color depths of the three tubes' imaging areas. In Table 5, the ratio of the average pixel values for the three tubes of solutions of C2 is 1:21.82:158.67. Even though the ratio of the average pixel values of the three tubes does not match the ratio of the permittivity values of the three solutions as shown in Table 2, the size pattern is consistent. The discrepancies between the color depths of the

TABLE 5 Average pixel value ratio of the three tube solutions for imaging results in Part III.

Phantom	C1	C2	C3
AVP_1: AVP_2: AVP_3	1:0.91:1.24	1:21.82:158.67	1:1.18:1.45

three tube locations in the image of C3 are minute. Table 5 provides the average pixel value ratio of 1:1.18:1.45 for the three tube regions in C3 imaging, which grows steadily although the difference is similarly minimal. Also, 1, 2, and 3 tubes in C3 are filled with alcohol, water, and blood solutions, respectively. The relative permittivity of alcohol and water is 40 and 80, respectively, and that of blood is somewhat more than that of water. The ratio of permittivity of the three solutions climbs progressively from tiny to large, mirroring the ratio of the three tubes' average pixel values. Nevertheless, the average pixel value ratio of the three tubes in C3 is significantly lower than that in C2, as evidenced by the images' color tones. Because vegetable oil has a permittivity of 2–4, the contrast between the permittivity of vegetable oil, alcohol, and blood in C2 is larger than that of alcohol, water, and blood in C3. The three-part physical experiment demonstrates that our developed ECT system is capable of imaging blood, with the pixel values reflecting the varied permittivity distributions.

4.4 Experimental results of brain hemorrhage imaging on isolated pigs

Figure 5 depicts the results of isolated porcine brain hemorrhage imaging tests. The red dashed circles in Figure 5 depict the real blood position. Although the imaging results can approximate the presence of blood, they deviate from the actual blood position. No matter whether blood was at 1/2 of the radius at the left or below in the actual situation, the imaging results deviated by about 3 mm from the center of the imaging area. In addition to the blood image area, there are other small regions with high pixel values in the images, notably around the electrodes. The most probable explanation for this is that the permittivity of blood is slightly larger than that of the isolated porcine brain and hence does not crush it. This effect is primarily attributable to the frozen porcine brain that was thawed and then doped with melt water, resulting in the porcine brain's permittivity being drastically lowered. Thus, the actual difference of permittivity between pig brain and blood is minimal. Second, the pig brain belongs to the non-uniform dielectric distribution, which contains gray matter, white matter, cerebrospinal fluid, and a small amount of residual blood, so it is in homogeneous medium, thus leading to a large difference between the sensitivity distribution of the imaging area full of pig brain and full of air, and the sensitivity matrices used for imaging in this paper are all calculated when the imaging zone is full of air, leading to poor imaging of blood, which may also be the reason why the imaging results deviate from the actual location. Nevertheless, the imaging results reveal the presence of blood more distinctly. Since the permittivity of the plastic cylinder material is very small and the permittivity of the actual skull is smaller than that of other brain tissues, the large cylinder wall can be regarded as the skull, so that the *ex vivo* experimental model can be approximately equivalent to the structure of an actual human head; thus, this *ex vivo* pig brain hemorrhage imaging experiment demonstrates the feasibility of the technique.

5 Conclusion

Extremely high morbidity and mortality rates are associated with ICH, and early detection and treatment are the keys to reducing mortality and enhancing postoperative results. In addition, currently, there is no portable and diminutive brain hemorrhage detection gadget. In this study, we constructed a 16-electrode ECT system based on an impedance analyzer and proved its viability for imaging brain hemorrhage by means of numerical simulation and physical measurements. In the simulation tests, a five-layer spherical brain hemorrhage model was created, and the imaging results precisely depicted the location and size of the hemorrhage. In physical studies, an isolated pig brain hemorrhage model was created and measured by the developed ECT system for differential imaging; the imaging findings similarly demonstrated the existence of brain hemorrhage, although the location of the hemorrhage in the image was somewhat altered relative to the actual position. In conclusion, the results of the simulation and the *ex vivo* imaging experiments confirmed the feasibility of ECT for brain hemorrhage imaging; however, the accuracy and resolution of the imaging were not high enough to be used for actual brain hemorrhage imaging. Therefore, improvements are required. The subsequent stage is to initially enhance the ECT system's performance. Second, a more advanced imaging algorithm should be developed to address the issue of imaging bias in *ex vivo* research. Additionally, ICH imaging tests *in vivo* should be conducted to examine the efficacy of ECT in actual hemorrhage imaging.

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

GJ: methodology and ECT system design; RX and WZ: hardware design, simulation experiments, and physical experiments; ZB: algorithmic design; GJ and NL: writing—reviewing and editing.

Funding

This research was funded by the Foundation of Scientific and Technological Innovation Capability Promotion of the Army Medical University (2019XQY06).

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

References

1. Collaborators GBDS, Global, regional, and national burden of stroke, 1990-2016: A systematic analysis for the global burden of disease study 2019. *Lancet Neurol* (2021) 18, 439–58.

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. AyazLzzetogluIzzetoglu HMK, 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. BodensteinDavidMarkstaller MMK. 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[J]. *Meas Sci Tech* (2001) 12:1126. doi:10.1088/0957-0233/12/8/319

7. Braun F, Proena M, Lemay M, Limitations and challenges of EIT-based monitoring of stroke volume and pulmonary artery pressure[J]. *Physiol Meas* (2018) 39(1):014003. IOP Publishing. doi:10.1088/1361-6579/aa9828

8. Toivanen J, Hnninen A, Savolainen T, Monitoring hemorrhagic strokes using EIT [J] (2021).

9. Mcdermott B, Elahi A, Santorelli A, O'Halloran M, Avery J, Porter E, Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis. *Physiol Meas* (2020) 41(7):075010 doi:10.1088/1361-6579/ab9e54

10. Watson S, Williams RJ, Griffiths H, A transceiver for direct phase measurement magnetic induction tomography[C]//International Conference of the IEEE Engineering in Medicine and Biology Society. In: Proceeding of the Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE. IEEE (2001).

11. Gabriel S, Lau RW, Gabriel C, 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(11):2251–69. doi:10.1088/0031-9155/41/11/002

12. Be Z, Ya Q, Li G, Detection of rabbit intracranial hemorrhage based on permittivity[J]. *Meas Sci Tech* (2019)(11) 30.

13. Bai Z, Li H, Chen J, Zhuang W, Li G, Jin G, 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

14. Rashid WNA, Johana Rahim E, AbdulAbdullah R, Jaafar A, Mohmad HL. Electrical capacitance tomography: A review on portable ECT system and hardware design[J]. *Sensor Rev* (2016) 36(1).

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

16. WangYang HW. Application of electrical capacitance tomography in pharmaceutical fluidised beds – a review. *Chem Eng Sci* (2020) 231:116236. doi:10.1016/j.ces.2020.116236

17. Taruno WP, A novel sensor design for breast cancer scanner based on electrical capacitance volume tomography (ECVT). *IEEE Sensors* (2012) 1 – 4.

18. Taruno WP, Brain tumor detection using electrical capacitance volume tomography. In: 6th International IEEE/EMBS Conference on Neural Engineering (NER) (2013). 743–6.

19. Taruno WP, Electrical Capacitance Volume Tomography for human brain motion activity observation. In: Proceeding of the Middle East Conference on Biomedical Engineering (MECBME) (2014). 147–50.

20. Yang W. Q., Peng L. Image reconstruction algorithms for electrical capacitance tomography[J]. *Meas Sci Tech* 2003(1):14.

21. Ye J, Wang H, Yang W. Image reconstruction for electrical capacitance tomography based on sparse representation[J]. *IEEE Trans Instrumentation Meas* (2014) 64(1):89–102.

22. Cui Z, Qi W, Qian X, Fan W, Zhang L, Yang W, et al. A review on image reconstruction algorithms for electrical capacitance/resistance tomography. *Sensor Rev* (2016) 36(4):429–45. doi:10.1108/sr-01-2016-0027

23. Yusuf A, Harry SS, Sudiana D, Tamsir AS, Widada W, Taruno WP, et al. Switch configuration effect on stray capacitance in electrical capacitance volume tomography hardware[J]. *TELKOMNIKA (Telecommunication Computing Electronics and Control)* (2016) 14(2):456. doi:10.12928/telkomnika.v14i2.3328

24. Axelsson O. *Iterative solution methods*. Cambridge: Cambridge University Press (1994).

25. Yu C, Han J, Song Y, Liu W A novel conjugate gradient image reconstruction algorithm for electrical capacitance tomography system[C]//. In: Proceeding of the International Conference on Challenges in Environmental Science and Computer Engineering. IEEE (2010).

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 06 April 2023
DOI 10.3389/fphy.2023.1180100

OPEN ACCESS

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

REVIEWED BY
Guanqiu Qi,
Buffalo State College, United States
Puhong Duan,
Hunan University, China

*CORRESPONDENCE
Yafei Zhang,
zyfeimail@kust.edu.cn

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 05 March 2023
ACCEPTED 23 March 2023
PUBLISHED 06 April 2023

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

Infrared and visible image fusion with edge detail implantation

Junyu Liu, Yafei Zhang* and Fan Li

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

Infrared and visible image fusion aims to integrate complementary information from the same scene images captured by different types of sensors into one image to obtain a fusion image with richer information. Recently, deep learning-based infrared and visible image fusion methods have been widely used. However, it is still a difficult problem how to maintain the edge detail information in the source images more effectively. To address this problem, we propose a novel infrared and visible image fusion method with edge detail implantation. The proposed method no longer improves the performance of edge details in the fused image through making the extracted features contain edge detail information like traditional methods, but by processing source image information and edge detail information separately, and supplementing edge details to the main framework. Technically, we propose a two-branch feature representation framework. One branch is used to directly extract features from the input source image, while the other is utilized to extract features of edge map. The edge detail branch mainly provides edge detail features for the source image input branch, ensuring that the output features contain rich edge detail information. In the fusion of multi-source features, we respectively fuse the source image features and the edge detail features, and use the fusion results of edge details to guide and enhance the fusion results of source image features so that they contain richer edge detail information. A large number of experimental results demonstrate the effectiveness of the proposed method.

KEYWORDS

infrared and visible image fusion, edge detail implantation, information compensation, dual branch network, end-to-end network

1 Introduction

Due to the different imaging mechanisms, two types of images for the same scene often carry a large amount of complementary information. If these complementary information can be integrated into one image, it will help improve the comprehensiveness and accuracy of the image to describe the scene, which is conducive to the development of subsequent tasks. To this end, infrared and visible image fusion technology has been proposed and widely applied to computer vision fields with different tasks, such as object detection [1], face recognition [2], video surveillance [3] and so on.

In recent years, with the rapid development of deep learning, the research of fusion methods among diverse modal information has made significant progress [4-8]. As an important branch in the field of image fusion, infrared and visible image fusion has attracted the attention of researchers, and a series of effective methods have been proposed. These methods can be roughly divided into methods based on multi-scale transformations [9-11], sparse modeling [12-15], and deep learning [16, 17]. Multi-scale transformation based methods include pyramid transform [9], DWT [18], Contourlet

transform (CT) [19], non-subsampled contourlet transform (NSCT) methods[20], etc. This kind of methods cannot achieve sparse expression of image because they use artificially constructed basis functions to represent images, limiting the visual quality improvement of fused images. Methods based on sparse modeling can solve the above problems by using an over-complete dictionary to represent images. However, these methods are difficult to mine the statistical information from large-scale training samples in an effective way, which limits the further improvement of their expression ability.

Among deep learning-based fusion methods, CNN-based methods are most common. At present, there are CNN-based infrared and visible image fusion as Cross-UNet-based [21], ResNet-based [22], GAN-based [23], Encoder-Decoder-based methods [24], etc. In view of the fact that CNNs cannot capture features over long distances, transformer-based infrared and visible image fusion method was proposed. However, since Transformer is designed based on attention mechanism, it has certain limitations in mining detailed information at the edges of the image. To solve the above problems, this paper proposes an infrared and visible image fusion method with edge detail implantation. In terms of feature extraction, the proposed method consists of two feature extraction branches: one is the feature extraction branch based on Transformer, and the other is the edge detail feature extraction branch based on CNN. The former takes infrared and visible source images as input, and the latter takes edge details detected from the source images as input. Information extracted by the latter is fed back to the former to compensate for the limitations of the transformer in extracting features.

In the feature implantation from CNN branch to the Transformer branch, an effective feature implantation method based on attention mechanism is designed, which not only considers the role of common information between different features in two branches, but also the complementary features extracted by CNN branch, realizing effective transmission of CNN features to Transformer branch. In terms of feature fusion, the features extracted from CNN branch and Transformer branch are fused respectively, and the fusion features of the CNN branch are used to guide the fusion features of the Transformer branch, so as to realize the fusion feature transfer from the CNN branch to the Transformer branch. It further compensates the shortcomings of the Transformer in feature extraction. The above method not only combine the advantages of CNN and Transformer in feature extraction into the whole framework, but also effectively enhances the representation ability of edge details, thereby improving the visual quality of the fusion results. In summary, the main contributions of this paper are as follows.

1) A method of infrared and visible image fusion with edge detail information implantation is proposed. This method uses two different branches based on Transformer and CNN to extract features from the input source images and the edge maps of the source images, and implants features extracted by CNN into the Transformer branch to make up for the shortcomings of Transformer in extracting edge details.
2) Based on attention mechanism, an information implantation method is designed, which realizes the injection of CNN branch information into Transformer, effectively making up for the shortcomings of Transformer in extracting features. In addition, the proposed method fuses the features obtained by CNN and Transformer branches respectively, and uses the fusion results of CNN branch to guide the fusion results of Transformer, further maintaining the edge details in fusion results.
3) In order to ensure that the features used to reconstruct fusion result are rich in edge details, we introduce an edge reconstruction block, and use edge detail information of the target image (fusion result) as a constraint to make the reconstruction result consistent with the target image, so as to ensure that feature to reconstruct the fusion result contains relevant information about edge details.

2 Related works

2.1 Infrared-visible image fusion

Infrared and visible image fusion is an important branch of fusion field. According to the previous introduction, current infrared and visible image fusion methods can be divided into fusion methods based on multi-scale transformation [25], sparse representation [26-29], and deep learning [30-32]. Multi-scale transformation based methods usually perform multi-scale decomposition of the input source image first, then fuse the decomposed coefficients, and finally apply the corresponding multi-scale inverse transformation to the fusion result to reconstruct the fusion image. These methods are simple to implement and has good stability. However, due to the use of fixed bases to represent the image information, their sparse expression ability is weak, which limits the improvement of fusion performance. Methods based on sparse representation were more popular 10 years ago. These methods can represent source images in a sparser way, obtaining better fusion performance than the former. However, these methods are difficult to mine statistical characteristics of features from large-scale training samples in an effective way, and thus still have limitations in representing image information.

In recent years, deep learning has been widely used in various image fusion tasks due to its powerful feature extraction and representation capabilities [33-38] without manually designing features and fusion strategies. In particular, Li et al. [31] proposed the DenseFuse infrared-visible image fusion framework, which combined the shallow and deep features of network by using dense blocks in the encoding process to extract richer source image features. In order to improve the fusion performance, Ma et al. [39] proposed a dual-confrontation DDcGAN fusion framework to further improve the performance of FusionGAN [23] when fusing infrared and visible images. Additionally, to more effectively maintain the edge details of the source image, Zhao et al. [40] used different encoders to extract high-frequency detail information and low-frequency information from the source images separately. Li et al. [22] proposed to use the detail preservation loss function and feature enhancement loss function in the residual structure network, combining with the two-stage training strategy to ensure that the fusion results contain rich detail information

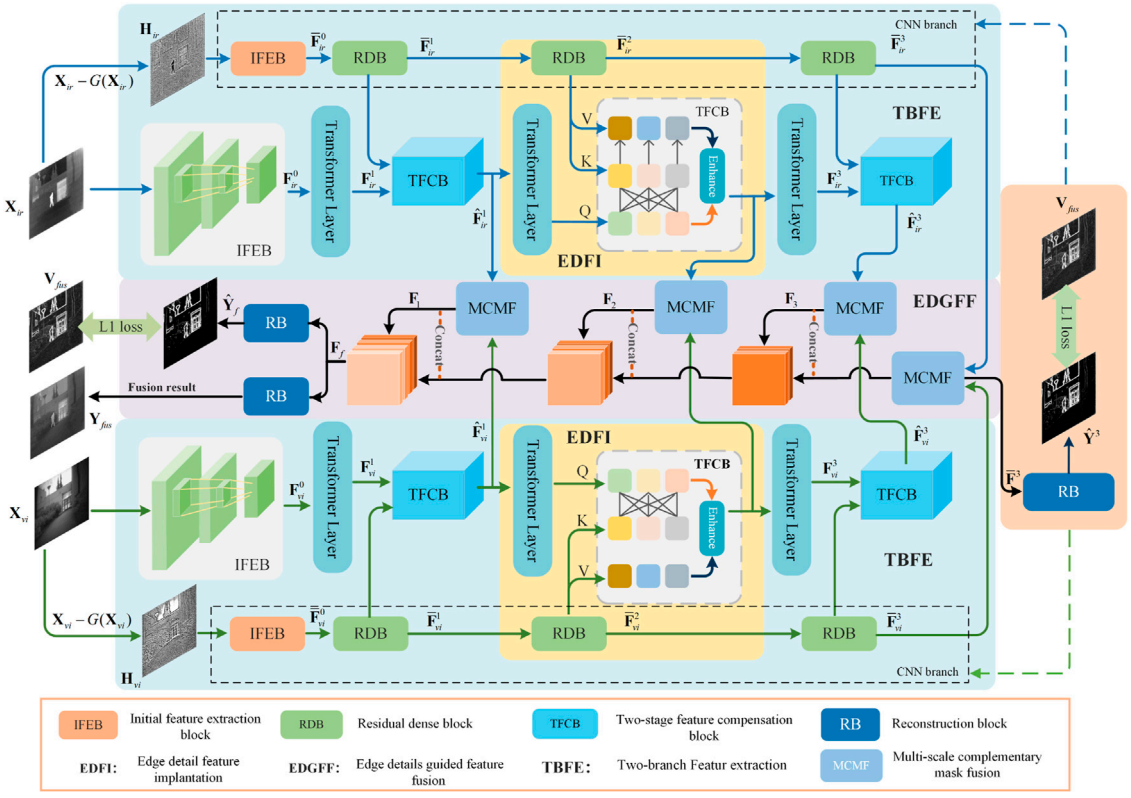


FIGURE 1
Overall framework of the proposed method.

and significant information. Although the above CNN-based methods have achieved certain performance, they are still insufficient in maintaining edge detail information. In addition, CNNs cannot mine the relationship between features over long distances, which limits the further improvement of their performance.

2.2 Transformer based image fusion

Thanks to its excellent long-distance modeling capabilities, Transformers [41] has attracted the attention of researchers in image fusion. In particular, in order to establish the global dependence of image features, Ma et al. [42] proposed a residual fusion framework based on SwinTransformer for infrared and visible image fusion. This framework abandons traditional convolution operations and adopts an attention-based network structure. At the same time, a fusion strategy based on L1 norm is designed, which further improves the fusion quality. In order to obtain high-quality pan-sharpened remote sensing images, Bandara et al. [43] proposed a new hyperTransformer framework. This method transfers the high-resolution texture information in PAN images to LR-HSI image features by attention mechanism, avoiding the image spatial and spectral distortion caused by traditional fusion methods.

In order to combine the respective advantages of CNN and Transformer, Vs. et al. [44] proposed a transformer-based encoding and decoding structure, and used the dual-branch structure of CNN and Transformer to fuse image features. Although this method can obtain satisfactory fusion results, it does not consider the problem of maintaining the edge details of source images, resulting in the loss of source image detail information. Li et al. [45] combined the local features of the convolutional network with the global features of the transformer by alternately using CNN and Transformer in the network, overcoming the shortcomings of a single network and improving the visual quality of the fused image. Based on the multi-scale feature pyramid theory, Park et al. [46] proposed an image fusion method for dual-modality transformers. This method mines the complementary information between source images by estimating the non-correlated mapping relationship between features of the source images, so as to improve the extracted feature quality of the source images. Although the above methods have achieved a certain degree of performance improvement, they does not fully consider the problem of maintaining the edge detail of the source images, which still remains large improvement space of visual effect. Different from the above methods, this paper uses two parallel feature extraction branches, Transformer and CNN, to extract the features of input source images and edge details, respectively, and implant the features extracted by CNN into the Transformer branch. This method can not only effectively integrate the respective advantages of CNN and Transformer, but also avoid the loss of edge detail information.

3 Proposed method

3.1 Overview

The overall framework of the proposed method is shown in Figure 1. It consists of three parts: two-branch feature extraction (TBFE), edge detail feature implantation (EDFI), and edge detail guided feature fusion (EDGFF). TBFE is mainly used to obtain the features and edge details of the source images. In this process, we use Transformer-based network to extract source image features, while utilize CNN network to the extract edge detail features. EDFI is mainly used to inject edge detail features extracted from the CNN branch into the Transformer branch to make up for the Transformer's shortcomings in extracting edge details. EDGFF uses the fused features of CNN branch to guide the fusion of Transformer branch, further highlighting the edge details in the fusion results.

3.2 Two-branch feature extraction

3.2.1 Transformer feature extraction branch

As shown in Figure 1, we use CNN and Transformer for TBFE, respectively, and the TBFEs for infrared and visible images have the same network structure. Since Transformer network has better feature relationship modeling ability for long-distance and can better describe the relationship between different features, we use Transformer branch to extract source image features. This branch takes the source image $\mathbf{X}_j$ ($j = ir, vi$) as the input, and first uses initial feature extraction block (IFEB) to obtain a shallow multi-channel feature map, which is convenient for subsequent Transformer feature extraction. The extracted features can be represented as:

$$\mathbf{F}_j^0 = f_{ifeb}\left(\mathbf{X}_j\right) \tag{1}$$

where f_{ifeb} denotes the feature extraction operation of IFEB. In this paper, IFEB consists of three 3×3 convolutional layers and a ReLU activation function. We utilize Transformer to extract the global features of the obtained feature $\mathbf{F}_j^0$. For the first transformer layer, its input feature is $\mathbf{F}_j^0$, and the output is expressed as:

$$\mathbf{F}_j^1 = f_{t1}\left(\mathbf{F}_j^0\right) \tag{2}$$

where f_{t1} represents the first Transformer layer feature extraction operation, mainly composed of layer normalization (LN), multi-head self-attention layer (MSA) and multi-layer perceptron (MLP). Correspondingly, for the ith ($i \geq 2$) Transformer layer, its output is $\mathbf{F}_j^i$.

3.2.2 CNN feature extraction branch

Compared with Transformer, CNNs are better at describing underlying visual features such as image structure and texture. Therefore, this paper uses CNN branch to extract features of edge details. In order to obtain the edge detail information from the source images, we perform Gaussian smoothing filtering on the source image $\mathbf{X}_j$ to obtain smooth image, and use the source image information to differ from the smooth image to obtain the edge detail information:

$$\mathbf{H}_j = \mathbf{X}_j - G\left(\mathbf{X}_j\right) \tag{3}$$

where G is the Gaussian blur operation. Edge details obtained in this way contain high-frequency information of the source image, which can effectively depict the edge details. Compared with the gradient map extracted by gradient operator, $\mathbf{H}_j$ contains richer edge detail and texture information. Similar to the Transformer branch, we use IFEB to extract the underlying features of the edge detail map in CNN branch:

$$\bar{\mathbf{F}}_j^0 = f_{ifeb}\left(\mathbf{H}_j\right) \tag{4}$$

Besides, detailed features are further extracted by residual dense block (RDB). For the first RDB, its input is features of the edge detail map $\bar{\mathbf{F}}_j^0$ and the output is $\bar{\mathbf{F}}_j^1$:

$$\bar{\mathbf{F}}_j^1 = f_{rdb1}\left(\bar{\mathbf{F}}_j^0\right) \tag{5}$$

where f_{rdb1} represents the feature extraction operation of the first RDB. In this work, RDB is a feature extraction block composed of three convolutional layers, ReLU activation function and densely connected between them. Correspondingly, for the ith ($i \geq 2$) RDB, its output is $\bar{\mathbf{F}}_j^i$.

3.3 Edge detail feature implantation

In order to make the extracted features in Transformer branch rich in edge detail, a two-stage feature compensation block (TFCB) is proposed, as shown in Figure 2. This module solves the problem that the Transformer branch is difficult to extract edge detail features by implanting local texture details of the image extracted by RDB into the Transformer branch. As for network structure, the module consists of two stages of feature compensation. The feature compensation in the first stage realizes the transmission of information by finding the correlation between $\bar{\mathbf{F}}_j^i$ and $\mathbf{F}_j^i$, and dynamically aggregating features of $\bar{\mathbf{F}}_j^i$ according to the changes of input features. Specifically, $\bar{\mathbf{F}}_j^i$ and $\mathbf{F}_j^i$ are first transformed into three feature spaces $\mathbf{Q}_j^i$, $\bar{\mathbf{K}}_j^i$, and $\bar{\mathbf{V}}_j^i$ by 1×1 convolution.

$$\begin{cases} \mathbf{Q}_j^i = \mathrm{Conv}_{1\times1}\left(\mathbf{F}_j^i\right) \\ \bar{\mathbf{K}}_j^i = \mathrm{Conv}_{1\times1}\left(\bar{\mathbf{F}}_j^i\right) \\ \bar{\mathbf{V}}_j^i = \mathrm{Conv}_{1\times1}\left(\bar{\mathbf{F}}_j^i\right) \end{cases} \tag{6}$$

where $\mathrm{Conv}_{1\times1}$ denotes 1×1 convolution. The first stage feature compensation process can be formulated as:

$$\tilde{\mathbf{F}}_j^i = \mathbf{F}_j^i + softmax\left(\frac{\mathbf{Q}_j^i\left(\bar{\mathbf{K}}_j^i\right)^T}{\sqrt{C}}\right)\bar{\mathbf{V}}_j^i \tag{7}$$

Where C is the dimension of $\bar{\mathbf{K}}_j^i$. The above method achieves information transfer from $\bar{\mathbf{F}}_j^i$ to $\mathbf{F}_j^i$ by using $\bar{\mathbf{F}}_j^i$ to represent $\mathbf{F}_j^i$. However, due to differences between $\bar{\mathbf{F}}_j^i$ and $\mathbf{F}_j^i$, the features re-aggregated based on similarity may lose some details. In order to avoid this problem, this paper introduces the second stage of feature compensation. Specifically, we input the re-aggregated features

$$\mathbf{T}_j^i = softmax\left(\frac{\mathbf{Q}_j^i\left(\bar{\mathbf{K}}_j^i\right)^T}{\sqrt{C}}\right)\bar{\mathbf{V}}_j^i \tag{8}$$

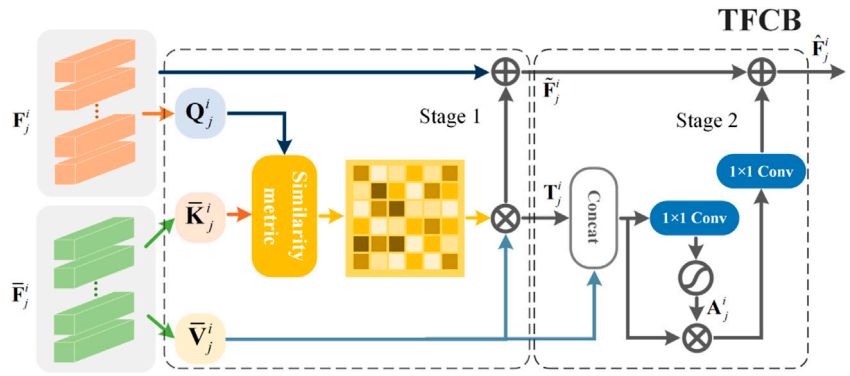


FIGURE 2
Detailed structure of TFCB.

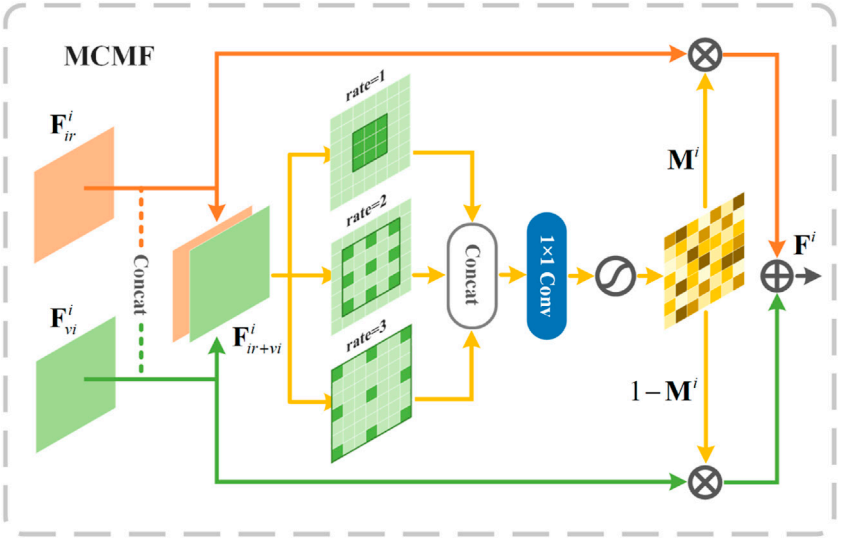


FIGURE 3
Detailed structure of MCMF.

and $\bar{\mathbf{V}}_j^i$ to a small CNN network, and select the activated features of the network through attention map, performing the second information compensation, as shown in Figure 2. To obtain the spatial attention map, we first concatenate $\mathbf{T}_j^i$ and $\bar{\mathbf{V}}_j^i$, and apply 1 × 1 convolution and Sigmoid:

$$\mathbf{A}_j^i = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{concat}\left(\mathbf{T}_j^i, \bar{\mathbf{V}}_j^i\right)\right)\right) \tag{9}$$

where σ represents the Sigmoid activation function. Output features of the ith information compensation module can be represented as:

$$\hat{\mathbf{F}}_j^i = \tilde{\mathbf{F}}_j^i + \mathrm{Conv}_{1\times1}\left(\mathbf{A}_j^i \odot \mathrm{concat}\left(\mathbf{T}_j^i, \bar{\mathbf{V}}_j^i\right)\right) \tag{10}$$

The two-stage feature compensation strategy not only avoids the shortcomings of Transformer in extracting edge detail features, but also prevents the loss of edge detail features and improves the quality of features, which helps to reconstruct high-quality fusion results.

3.4 Edge details guided feature fusion

In order to effectively use the edge detail features of the CNN branch to supplement the features in the Transformer branch in fusion image reconstruction, this paper proposes a method to synthesize the edge detail fusion results of the CNN branch and the fusion results of the Transformer branch to jointly construct the final fusion results. In order to effectively fuse the multimodal features extracted by TBFE. We design a multi-scale complementary mask fusion (MCMF) module to ensure its effectiveness. As shown in Figure 3, MCMF concatenate the features $\mathbf{F}_{ir}^i$ and $\mathbf{F}_{vi}^i$ from the output of the Transformer branch to obtain $\mathbf{F}_{ir+vi}^i$, which is feed into the convolutional layer to learn the weight map $\mathbf{M}^i$ for fusion. In this process, we apply three dilated convolutions with different dilation rates to the concatenated features, mining the importance information in different receptive fields in a

more flexible way. After concatenating three groups of results, the feature fusion is performed by 1 × 1 convolution, and the fusion weight map that reflects the importance of each position in the source image features is obtained through the Sigmoid activation function.

$$\mathbf{M}^i = \mathrm{Sigmoid}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{concat}\left(\mathrm{Conv}_{3\times3}\left(\mathbf{F}^i_{ir+vi}, \quad r=1\right),\right.\right.\right.$$
$$\left.\left.\left.\mathrm{Conv}_{3\times3}\left(\mathbf{F}^i_{ir+vi}, \quad r=2\right), \mathrm{Conv}_{3\times3}\left(\mathbf{F}^i_{ir+vi}, \quad r=3\right)\right)\right)\right) \tag{11}$$

where $\mathrm{Conv}_{3\times3}$ denotes 3 × 3 convolution and r is dilation rate.

The fusion feature can be expressed as:

$$\mathbf{F}^i = \mathbf{F}^i_{ir} \odot \mathbf{M}^i + \mathbf{F}^i_{vi} \odot \left(1 - \mathbf{M}^i\right) \tag{12}$$

where ⊙ denotes hadamard product. Similar to the fusion of Transformer features, the edge detail features $\bar{\mathbf{F}}^i_{ir}$ and $\bar{\mathbf{F}}^i_{vi}$ extracted from the CNN branch are also fused in the above way to obtain the fusion result $\bar{\mathbf{F}}^3$ of the last RDB output features. The fused detailed features $\bar{\mathbf{F}}^3$ are concatenated with the fusion features of Transformer branch at three scales $\mathbf{F}^1$, $\mathbf{F}^2$, $\mathbf{F}^3$ to obtain $\mathbf{F}_f$. In order to ensure that both $\mathbf{F}_f$ and $\bar{\mathbf{F}}^3$ contain rich edge detail features, We reconstruct the edge detail feature maps by a reconstruction block (RB) for $\mathbf{F}_f$ and $\bar{\mathbf{F}}^3$ respectively, and make the reconstructed results consistent with the target feature maps. The RB used for reconstruction in this work consists of two 3 × 3 and one 1 × 1 convolutional layers, and the parameters are not shared between the different reconstruction block. Besides, this work uses the gradient detection operator to directly extract the gradient information of the source images, and fuse them to obtain high-quality target feature map $\mathbf{V}_{fus}$. The specific process is as follows:

$$\mathbf{V}_{fus}(i,j) = \begin{cases} \nabla\mathbf{X}_{ir}(i,j), if\,|\nabla\mathbf{X}_{ir}(i,j)| \geq |\nabla\mathbf{X}_{vi}(i,j)| \\ \nabla\mathbf{X}_{vi}(i,j), otherwise \end{cases} \tag{13}$$

where $\mathbf{V}_{fus}$ is the edge detail of the target image, ∇ is the Laplace operator, and (i, j) is the pixel coordinates.

4 Loss function

To ensure high visual quality fusion results, we use the L1 loss shown in Eq. 14 to optimize the parameters in the RDB:

$$\ell_f = \left\|\hat{\mathbf{Y}}^3 - \mathbf{V}_{fus}\right\|_1 \tag{14}$$

where $\hat{\mathbf{Y}}^3$ is the image after $\bar{\mathbf{F}}^3$ is reconstructed by RB. In the reconstruction process of the fusion results, the fusion image and the target feature map are reconstructed respectively with pixel loss, which is designed to limit the difference in intensity between real-world data and the reconstructed model result. $\mathbf{Y}_{fus}$ is the final fused image, and the reconstruction loss used in this work is as follows:

$$\ell_r = \|\mathbf{Y}_{fus} - \mathbf{X}_{ir}\|_1 + \|\mathbf{Y}_{fus} - \mathbf{X}_{vi}\|_1 + \left\|\hat{\mathbf{Y}}_f - \mathbf{V}_{fus}\right\|_1 \tag{15}$$

where $\hat{\mathbf{Y}}_f$ is the reconstructed image of $\mathbf{F}_f$ by RB. The total loss of the network is expressed as:

$$\ell_{total} = \ell_f + \ell_r \tag{16}$$

5 Experiments

5.1 Dataset

KAIST[1] and FLIR[2] are the two most commonly used datasets in the field of infrared and visible image fusion based on deep learning. Among them, there are 95,000 infrared and visible image pairs in the KAIST dataset and 14,452 image pairs in the FLIR dataset. In order to improve the generalization ability of the training model, 3,000 image pairs are randomly selected from the two datasets respectively, and a total of 6,000 image pairs from the training set of the proposed algorithm in this work. To verify the effectiveness of the method, 49 pairs of widely used infrared and visible images are randomly selected from the three datasets TNO[3], VOT2020-RGBT[4] and RoadScence[5] to construct the test set in this work. Among them, 39 pairs of images are from TNO and VOT2020-RGBT datasets and 10 pairs of images are from the RoadScence dataset. The test samples are shown in Figure 4.

5.2 Training details

In the training phase, each infrared and visible image pair is randomly cropped into 140 × 140 image blocks to achieve data enhancement. In this work, Adam [47] is used as the optimizer of the network, the training batchsize is set to 4, and a total of 30 epochs are iterated. The initial value of learning rate is set to 1×10^{-4}, and decays at the 5-th, 10-th, and 20-th epochs, respectively, with a decay rate of 0.5. The code of our method is implemented by using the PyTorch framework with NVIDIA GTX 3090, and the software environment is UBUNTU20.2, Python3.8 and PyTorch1.9.

5.3 Evaluation metrics

In order to objectively evaluate the fusion performance, six commonly used image fusion metrics are used in this work to assess the quality of fusion results from four perspectives. They are cross entropy (Q_{CE}) [48]; Entropy (Q_{EN}) [49]; gradient-based fusion performance ($Q_{AB/F}$) [50]; Chen -Blum metric (Q_{CB}) [51],; Chen-Varshney metric (Q_{CV}) [52]and Structural similarity index measure (Q_{SSIM}) [53]. Among them, Q_{CE} and Q_{EN} are metrics based on information theory ($Q_{AB/F}$) is a metrics based on image features, Q_{CB} and Q_{CV} are metrics based on human perception, and Q_{SSIM} is a metrics based on structural similarity of images. Among these six metrics, lower values for Q_{CE} and Q_{CV} indicate better quality of

1 https://soonminhwang.github.io/rgbt-ped-detection/.

2 https://www.flir.ca/oem/adas/adas-dataset-form/.

3 https://figshare.com/articles/dataset/TNO-Image-Fusion-Dataset/1008029.

4 https://www.votchallenge.net/vot2020/dataset.

5 https://github.com/hanna-xu/RoadScene.

FIGURE 4
Some test images from TNO, VOT2020-RGBT and RoadScence datasets.

TABLE 1 Quantitative evaluation of different fusion methods on 39 pairs of images from TNO and VOT2020-RGBT datasets. The red font indicates the optimal results and the blue font indicates the sub-optimal results.

Methods	Q_{CE} ↓	Q_{EN} ↑	$Q_{AB/F}$ ↑	Q_{CB} ↑	Q_{CV} ↓	Q_{SSIM} ↑
ADF	1.6913	6.4384	0.4021	0.4812	609.7199	1.4246
GTF	1.0113	6.7083	0.3346	0.4220	1,237.5143	1.4101
LatLRR	2.5355	6.7744	0.3596	0.4767	656.1416	1.1879
FusionGAN	2.2057	6.5336	0.2299	0.4045	1,025.8828	1.3489
SDNet	1.7178	6.6736	0.4527	0.4650	825.2304	1.4324
RFN	1.7270	6.8364	0.3602	0.4723	642.0348	1.4226
Ours	1.5498	6.9070	0.4815	0.5068	564.0687	1.4476

fusion results, while higher values for the other indicators indicate better fusion performance.

5.4 Comparison with state-of-the-arts

In order to verify the effectiveness of the proposed method, we compare our method with six advanced infrared and visible image fusion methods, including ADF [54], GTF [55], LatLRR [56], FusionGAN [23], SDNet [57], and RFN [22].

Figure 5 shows the fusion results of different methods on six groups of test images. In order to facilitate the fusion quality evaluation from the perspective of visual effect, we zoom in local area of the fusion results. It can be seen that the proposed method can not only preserve the salient information in the infrared image, but also maintain the edge detail information in the visible image. In detail, the outline of the infrared salient information is blurred, and

FIGURE 5
Fusion results of different methods on six pairs of images from TNO and VOT2020-RGBT datasets.

TABLE 2 Quantitative evaluation of different fusion methods on 10 pairs of images from RoadScence datasets. The red font indicates the optimal results and the blue font indicates the sub-optimal results.

Methods	Q_{CE} ↓	Q_{EN} ↑	$Q_{AB/F}$ ↑	Q_{CB} ↑	Q_{CV} ↓	Q_{SSIM} ↑
ADF	1.5202	6.7251	0.4084	0.4398	954.7853	1.3249
GTF	0.5876	7.0637	0.3708	0.4211	1964.4652	1.3212
LatLRR	2.0122	5.8881	0.3943	0.4285	880.5359	1.2419
FusionGAN	1.6016	6.9790	0.2939	0.4204	1,546.3186	1.2255
SDNet	1.6867	7.0493	0.4497	0.4604	1,587.5962	1.3285
RFN	1.3567	6.8826	0.3585	0.4390	1,398.5784	1.3139
Ours	1.5661	7.1463	0.4634	0.4839	741.7628	1.3601

the edge detail is not preserved enough in the fusion results obtained by FusionGAN, RNF, and ADF. In contrast, other methods effectively retain significant information, but the loss of spatial detail is more pronounced, as shown in the zoomed-in area of GTF fusion results. Similar phenomena can be observed in the remaining other fusion results. On the whole, the proposed method can more fully retain the significant edge detail information of the source images, and show better fusion performance. As for objective metrics, the proposed method has achieved excellent performance on Q_{EN}, $Q_{AB/F}$, Q_{CB}, Q_{CV} and Q_{SSIM}, as shown in Table 1, which further verifies the effectiveness of our method.

To further verify the effectiveness of the proposed method, we deploy the above comparison methods to the test data selected from the Roadscene. Figure 6 shows the fusion results of different

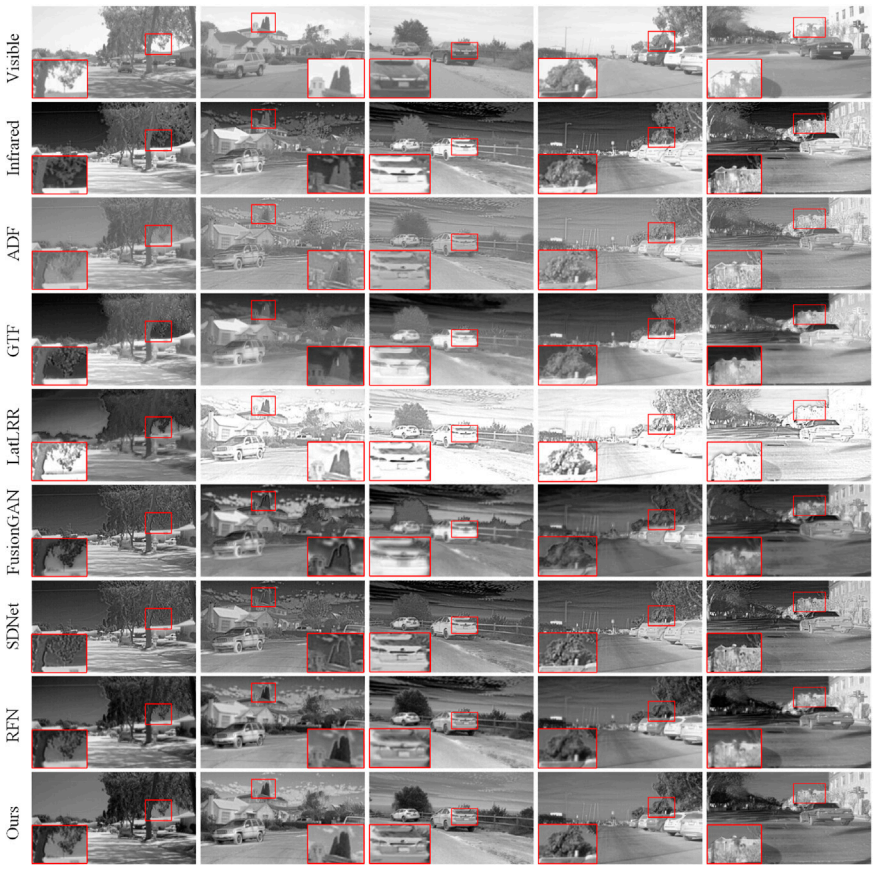


FIGURE 6
Fusion results of different methods on five pairs of images from RoadScence.

TABLE 3 Analysis of the effectiveness of different functional modules.

Methods	$Q_{CE}\downarrow$	$Q_{EN}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CB}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
w/o $\mathbf{H}_j$	1.5721	6.8866	0.4592	0.4857	600.7409	1.4166
w/o TFCB	1.5911	6.9014	0.4762	0.4780	569.5434	1.4244
w/o MCMF	1.5593	6.8715	0.4754	0.4812	588.4187	1.4406
w/o $\mathbf{V}_{fus}$	1.5557	6.8911	0.4547	0.4828	689.1718	1.4122
Ours	1.5498	6.9070	0.4815	0.5068	564.0687	1.4476

TABLE 4 The effect of different nuber of TFCBs on fusion performance.

Methods	$Q_{CE}\downarrow$	$Q_{EN}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CB}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
2 TFCBs	1.5687	6.8988	0.4805	0.4982	565.9757	1.4415
3 TFCBs	1.5498	6.9070	0.4815	0.5086	564.0687	1.4476
4 TFCBs	1.5425	6.8697	0.4805	0.4866	573.2399	1.4488

fusion methods on five pairs of test images from Roadscene. Deep learning-based fusion methods (FusionGAN, SDNet, and RFN) have better visual performance than traditional fusion methods (ADF, GTF, and LatLRR). Traditional fusion methods are limited by hand-designed fusion rules, resulting in problems such as too bright, too dark, or loss of detail information. FusionGAN adopts the adversarial learning network structure and lacks constraints on spatial consistency, causing blurred edge details in the obtained fusion results, which affects the visual quality improvement of the fusion images. SDNet, considers the spatial gradient information in network design, so it has certain advantages in detail retention, but its ability of remaining texture information is weak. Similar problem exists in RFN., in contrast, our fusion results have two advantages. First, the significant information of the infrared image can be well retained, so that the fusion results can further highlight the target, which is conducive to subsequent tasks (such as target detection, instance segmentation, etc.). Second, more texture detail information can be retained, ensuring the quality of the fusion results to a certain extent. In order to evaluate the quality of the fused images more comprehensively, we use six commonly used objective evaluation metrics to evaluate the quality of the fused images. From Table 2, results of the proposed method reach the optimal on five indicators of Q_{EN}, $Q_{AB/F}$, Q_{CB}, Q_{CV}, and Q_{SSIM}, which further proves the effectiveness and superiority of the proposed method.

5.5 Ablation study

In order to verify the influence of different components on the fusion performance of the proposed method, we apply ablation experiments on each module.

In the validation of the input of edge detail information, we use the source images instead of edge details as the input of CNN branch to verify the influence of edge details on fusion results (**w/o** $\mathbf{H}_j$). In the implantation of edge detail feature, we use two-stage feature compensation block (TFCB) to compensate the information of the Transformer branch. In order to verify the effectiveness of edge detail feature implantation, we add the features of two branch to replace TFCB module (**w/o** TFCB). In edge detail-guided feature fusion, the multi-scale complementary mask fusion (MCMF) module is the key. To verify its effectiveness, MCMF is replaced by conventional feature channel concatenation and 1 × 1 convolution to achieve feature fusion (**w/o** MCMF). The target edge detail map $\mathbf{V}_{fus}$ is used to enrich the features of the reconstructed fusion result with edge detail information. To demonstrate its validity, we directly removes it from the model (**w/o** $\mathbf{V}_{fus}$). The effectiveness of the above modules is tested on 39 pairs of images from TNO and VOT2020-RGBT. The effectiveness of the different components can be seen in Table 3.

5.6 Hyper-parameter analysis

The number of TFCBs determines the depth of the network, whose impact on the final performance can be seen in Table 4. When the number of TFCBs modules is three, the optimal result is obtained among the six evaluation indexes overall, so we set the number of TFCBs to three.

6 Conclusion

In order to effectively maintain the edge detail information of the source images, we propose a infrared and visible image fusion method with edge detail implantation, which adopts a two-branch feature representation framework. One branch is based on Transformer, which is mainly used to directly extract features from input source images. The other is CNN feature extraction branch, which is mainly used to extract image edge details features. Features extracted by CNN branch are implanted into the Transformer branch to alleviate the shortcomings of the Transformer branch in extracting edge detail features. In addition, so as to further ensure that the edge details of the source image can be effectively retained in the fusion results, a feature fusion method guided by edge details is proposed, which uses the fused edge detail features of CNN branch to guide the feature fusion of Transformer branch. A large number of experimental results prove the effectiveness of our method.

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

JL responsible for paper scheme design, experiment and paper writing. YZ guide the paper scheme design and revision. FL guide to do experiments and write papers. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 62161015). Name of Fund: Research on multi-source image fusion algorithm unconstrained by registration and resolution consistency.

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. Han J, Bhanu B. Fusion of color and infrared video for moving human detection. *Pattern Recognition* (2007) 40:1771–84. doi:10.1016/j.patcog.2006.11.010

2. Singh R, Vatsa M, Noore A. Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition. *Pattern Recognition* (2008) 41:880–93. doi:10.1016/j.patcog.2007.06.022

3. Kumar P, Mittal A, Kumar P. Fusion of thermal infrared and visible spectrum video for robust surveillance. In: Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006; December 13-16, 2006; Madurai, India (2006). p. 528–39.

4. Tang L, Deng Y, Ma Y, Huang J, Ma J. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA J Automatica Sinica* (2022) 9:2121–37. doi:10.1109/jas.2022.106082

5. Liu Y, Shi Y, Mu F, Cheng J, Li C, Chen X. Multimodal mri volumetric data fusion with convolutional neural networks. *IEEE Trans Instrumentation Meas* (2022) 71:1–15. doi:10.1109/tim.2022.3184360

6. Liu Y, Wang Z. Dense sift for ghost-free multi-exposure fusion. *J Vis Commun Image Representation* (2015) 31:208–24. doi:10.1016/j.jvcir.2015.06.021

7. 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

8. 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

9. Vanmali AV, Gadre VM. Visible and nir image fusion using weight-map-guided laplacian–Gaussian pyramid for improving scene visibility. *Sādhanā* (2017) 42:1063–82. doi:10.1007/s12046-017-0673-1

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 Sensors J* (2020) 20: 8062–72. doi:10.1109/jsen.2020.2981719

11. 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

12. Mitianoudis N, Antonopoulos SA, Stathaki T. Region-based ica image fusion using textural information. In: 2013 18th International Conference on Digital Signal Processing (DSP) (IEEE); Jul. 01 - 03, 2013; Greece (2013). p. 1–6.

13. Li H, He X, Tao D, 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

14. 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

15. Li H, He X, Yu Z, Luo J. Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior. *Inf Sci* (2020) 523:14–37. doi:10.1016/j.ins.2020.03.009

16. Xu H, Ma J, Jiang J, Guo X, U2fusion LH. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans Pattern Anal Machine Intelligence* (2020) 44:502–18. doi:10.1109/TPAMI.2020.3012548

17. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

18. Zhan L, Zhuang Y, Huang L. Infrared and visible images fusion method based on discrete wavelet transform. *J Comput* (2017) 28:57–71.

19. Yang B, Li S, Sun F. Image fusion using nonsubsampled contourlet transform. In: Fourth International Conference on Image and Graphics (ICIG 2007) (IEEE); 22-24 August 2007; Chengdu, China (2007). p. 719–24.

20. Li H, Qiu H, Yu Z, Zhang Y. Infrared and visible image fusion scheme based on nsct and low-level visual features. *Infrared Phys Tech* (2016) 76:174–84. doi:10.1016/j.infrared.2016.02.005

21. Wang X, Hua Z, Li J. Cross-UNet: Dual-branch infrared and visible image fusion framework based on cross-convolution and attention mechanism. *Vis Comput* (2022). doi:10.1007/s00371-022-02628-6

22. Li H, Wu XJ, Kittler J. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Inf Fusion* (2021) 73:72–86. doi:10.1016/j.inffus.2021.02.023

23. Ma J, Yu W, Liang P, Li C, Fusiongan JJ. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf fusion* (2019) 48:11–26. doi:10.1016/j.inffus.2018.09.004

24. Zhao Z, Xu S, Zhang C, Liu J, Li P, Zhang J. *Didfuse: Deep image decomposition for infrared and visible image fusion* (2020). *arXiv preprint arXiv:2003.09210*.

25. Liu Y, Liu S, Wang Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf fusion* (2015) 24:147–64. doi:10.1016/j.inffus.2014.09.004

26. Gao Y, Ma J, Yuille AL. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. *IEEE Trans Image Process* (2017) 26:2545–60. doi:10.1109/tip.2017.2675341

27. Li H, Wang Y, 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* (2019) 69:1082–102. doi:10.1109/tim.2019.2912239

28. Zhu Z, Yin H, 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

29. Yin M, Duan P, Liu W, Liang X. A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation. *Neurocomputing* (2017) 226:182–91. doi:10.1016/j.neucom.2016.11.051

30. Mo Y, Kang X, Duan P, Sun B, Li S. Attribute filter based infrared and visible image fusion. *Inf Fusion* (2021) 75:41–54. doi:10.1016/j.inffus.2021.04.005

31. Li H, Wu XJ. Densefuse: A fusion approach to infrared and visible images. *IEEE Trans Image Process* (2018) 28:2614–23. doi:10.1109/tip.2018.2887342

32. 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.doi:10.1016/j.inffus.2023.02.011

33. 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

34. 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 Instrumentation Meas* (2020) 70:1–23. doi:10.1109/tim.2020.3024335

35. Ma J, Zhao J, Jiang J, Zhou H, Guo X. Locality preserving matching. *Int J Comp Vis* (2019) 127:512–31. doi:10.1007/s11263-018-1117-z

36. 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

37. Zhu Z, Liang H, Li Y, Qi G. A method for quality evaluation of multi-exposure fusion images with multi-scale gradient magnitude. In: Proceedings of 2021 Chinese Intelligent Systems Conference: Volume II; Nov 5-7, 2021; Zhanjiang, China (2022). p. 121–9.

38. Liu Y, Wang L, Li H, Chen X. Multi-focus image fusion with deep residual learning and focus property detection. *Inf Fusion* (2022) 86:1–16. doi:10.1016/j.inffus.2022.06.001

39. Ma J, Xu H, Jiang J, Mei X, Zhang XP. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans Image Process* (2020) 29:4980–95. doi:10.1109/tip.2020.2977573

40. Zhao Z, Xu S, Zhang J, Liang C, Zhang C, Liu J. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans Circuits Syst Video Tech* (2022) 32:1186–96. doi:10.1109/tcsvt.2021.3075745

41. 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.

42. Ma J, Tang L, 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) 9:1200–17. doi:10.1109/jas.2022.105686

43. Bandara WGC, Patel VM. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 19-25 June 2021 (2022) 1767–77.

44. Vs V, Valanarasu JMJ, Oza P, Patel VM. Image fusion transformer. In: 2022 IEEE International Conference on Image Processing (ICIP) (IEEE); October 16-19, 2022; Bordeaux France (2022). p. 3566–70.

45. Li J, Zhu J, Li C, Chen X, Yang B. Cgtf: Convolution-guided transformer for infrared and visible image fusion. *IEEE Trans Instrumentation Meas* (2022) 71:1–14. doi:10.1109/tim.2022.3175055

46. Park S, Vien AG, Lee C. Infrared and visible image fusion using bimodal transformers. In: 2022 IEEE International Conference on Image Processing (ICIP) (IEEE); October 16-19, 2022; Bordeaux France (2022). p. 1741–5.

47. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations; April 14-16, 2014; Banff, AB, Canada (2014). p. 109–19.

48. Bulanon D, Burks T, Alchanatis V. Image fusion of visible and thermal images for fruit detection. *Biosyst Eng* (2009) 103:12–22. doi:10.1016/j.biosystemseng.2009.02.009

49. Roberts JW, Van Aardt JA, Ahmed FB. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J Appl Remote Sensing* (2008) 2:023522. doi:10.1117/1.2945910

50. Xydeas CS, Petrovic V. Objective image fusion performance measure. *Elect Lett* (2000) 36:308–9. doi:10.1049/el:20000267

51. Chen Y, Blum RS. A new automated quality assessment algorithm for image fusion. *Image Vis Comput* (2009) 27:1421–32. doi:10.1016/j.imavis.2007.12.002

52. Chen H, Varshney PK. A human perception inspired quality metric for image fusion based on regional information. *Inf fusion* (2007) 8:193–207. doi:10.1016/j.inffus.2005.10.001

53. 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

54. Bavirisetti DP, Dhuli R. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform. *IEEE Sensors J* (2015) 16:203–9. doi:10.1109/jsen.2015.2478655

55. Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf Fusion* (2016) 31:100–9. doi:10.1016/j.inffus.2016.02.001

56. Li H, Wu XJ. *Infrared and visible image fusion using latent low-rank representation* (2018). *arXiv preprint arXiv:1804.08992*.

57. Zhang H, Ma J. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int J Comp Vis* (2021) 129:2761–85. doi:10.1007/s11263-021-01501-8

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 28 April 2023
DOI 10.3389/fphy.2023.1194573

OPEN ACCESS

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

REVIEWED BY
Hui Li,
Jiangnan University, China
Wei Huang,
Zhengzhou University of Light Industry, China
Qingbei Guo,
University of Jinan, China

*CORRESPONDENCE
Zhenqiu Shu,
shuzhenqiu@163.com

RECEIVED 27 March 2023
ACCEPTED 17 April 2023
PUBLISHED 28 April 2023

CITATION
Li G, Peng Q, Zou D, Yang J and Shu Z (2023), Fine-grained similarity semantic preserving deep hashing for cross-modal retrieval.
Front. Phys. 11:1194573.
doi: 10.3389/fphy.2023.1194573

Fine-grained similarity semantic preserving deep hashing for cross-modal retrieval

Guoyou Li[1], Qingjun Peng[2], Dexu Zou[2], Jinyue Yang[2] and Zhenqiu Shu[3]*

[1]Yunnan Power Grid Corporation, Kunming, China, [2]Electric Power Research Institute, Yunnan Power Grid Corporation, Kunming, China, [3]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

Cross-modal hashing methods have received wide attention in cross-modal retrieval owing to their advantages in computational efficiency and storage cost. However, most existing deep cross-modal hashing methods cannot employ both intra-modal and inter-modal similarities to guide the learning of hash codes and ignore the quantization loss of hash codes, simultaneously. To solve the above problems, we propose a fine-grained similarity semantic preserving deep hashing (FSSPDH) for cross-modal retrieval. Firstly, this proposed method learns different hash codes for different modalities to preserve the intrinsic property of each modality. Secondly, the fine-grained similarity matrix is constructed by using labels and data features, which not only maintains the similarity between and within modalities. In addition, quantization loss is used to learn hash codes and thus effectively reduce information loss caused during the quantization procedure. A large number of experiments on three public datasets demonstrate the advantage of the proposed FSSPDH method.

KEYWORDS

cross-modal fusion, similarity semantic preserving, quantization loss, deep hashing, intra-modal similarity, inter-modal similarity, fine-grained similarity

1 Introduction

As electronic technology and the Internet have advanced, the amount of multimedia data, such as images, texts, audio, and video, has experienced rapid growth. Therefore, how to effectively implement cross-modal retrieval has become a hot research field. However, due to the differences in data distribution and feature representation between different modalities, it is a huge challenge in cross-modal retrieval to narrow the semantic gap between multimodal data. Generally, the goal of cross-modal retrieval is to map the original data into a common potential space to maintain the similarity structure of the original features and find the most similar samples in the new feature space [1]. In addition, hashing technology can significantly reduce storage space and computational complexity because it only requires binary operation. Therefore, it becomes an effective way to solve cross-modal retrieval of massive data [2–4].

Cross-modal hashing is generally divided into two main categories, which are supervised hashing and unsupervised hashing. The unsupervised hashing [5,6] aims to project data features into a common feature space to reduce the difference between modalities. The supervised hashing methods [7,8] use label information to further enhance the semantic correlation between cross-modal data. The use of label information significantly narrows the gap between modalities and achieves excellent retrieval performance. Since deep learning has demonstrated its strong

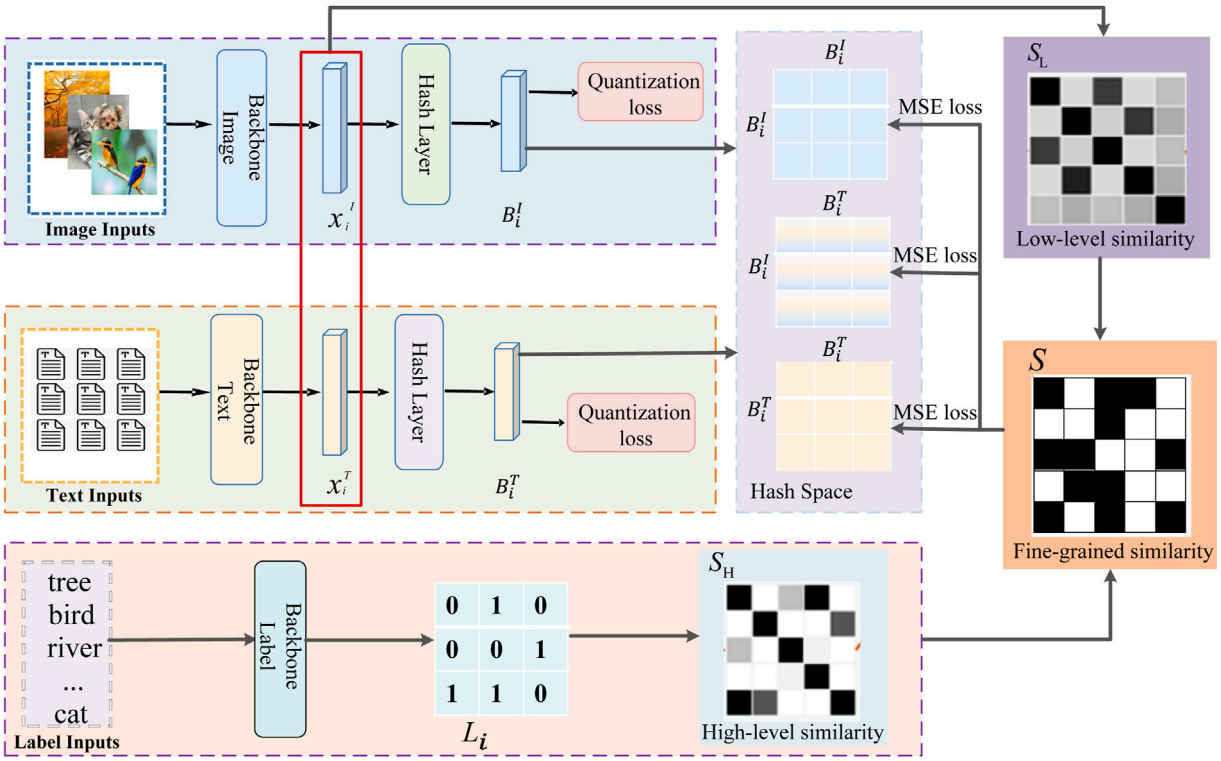


FIGURE 1
The framework of the proposed FSSPDH approach.

advantage in various fields, many deep cross-modal hashing methods have been proposed in recent years [9–11]. Tu et al. [12] proposed an end-to-end deep cross-modal hashing method, which obtains the unified hash codes of the training and the query samples through the joint learning of hash codes and hash functions. Self-supervised adversarial hashing (SSAH) [13] adopts two adversarial networks to jointly model semantic features of different modalities and then utilizes their semantic correlations to generate binary hash codes. At present, deep hashing methods have achieved excellent performances in cross-modal retrieve tasks, but there are still some issues to be solved urgently: 1) Most existing deep cross-modal hashing methods ignore the intra-modal and inter-modal similarities to guidance the hash code learning; 2) Existing hashing methods mainly focus on the hash code generation stage, and thus hash representations with less semantic information and spatial correlation cannot generate optimal hash codes; 3) Many methods often fail to consider the quantization loss of hash codes, resulting in the loss of semantic information during hash code learning.

To solve the above problems, we propose a fine-grained similarity semantic preserving deep hashing (FSSPDH) method for cross-modal retrieval tasks. Figure 1 shows the framework of our proposed FSSPDH method. The main contributions of this work are given as follows.

1) The proposed FSSPDH approach unifies data feature extraction and hash code learning into an end-to-end deep learning framework. It can learn different hash codes from different modalities and thus maintains the intrinsic property of each modality. In addition, the proposed method combines the high-level semantic similarity constructed with labels and the low-level semantic similarity constructed with features to construct a fine-grained similarity matrix. Compared with traditional similarity constraints, the fine-grained similarity can effectively maintain inter-modal and intra-modal similarities to explore the semantic relationship between modalities and instances.
2) Our FSSPDH method considers the quantization loss in hash code learning, which further reduces the information loss caused by the hash code quantization. The quantization loss can make the learned hash codes with more feature information obtain more discriminative hash codes.
3) Experimental results conducted on three widely used multimodal datasets indicate that our proposed FSSPDH method achieves higher accuracy in cross-modal retrieval tasks compared with other hashing methods.

The remaining parts of this paper are organized as follows: Section 2 reviews the related works of cross-modal hashing retrieval. In Section 3, we introduce our FSSPDH approach in detail. Section 4 describes the experimental results and their results. Finally, our work is drawn in Section 5.

2 Related work

At present, cross-modal hashing can be roughly divided into the unsupervised method and supervised method according to whether it uses supervised information. This section will give a brief overview of these two types of methods.

2.1 Unsupervised cross-modal hashing

Since most multimodal data from real life are unlabeled, it is unrealistic to consume significant labor and time to label these data. Therefore, unsupervised hashing methods have received extensive attention in cross-modal retrieval. These methods attempt to learn the correlation and underlying structure of multimodality data. They can be further divided into graph-based methods and matrix factorization-based methods. The former seeks to maintain the correlation of hash codes by constructing a similarity graph. Linear cross-modal hashing (LCMH) [14] uses an anchor graph to keep the similarity within and between models in Hamming space. Hetero-manifold regularisation (HMR) [15] constructs multiple sub-manifolds defined by homogeneous data with the help of supervision information and alleviates the integration complexity and heterogeneity problems. Fusion similarity hashing (FSH) [16] constructs an asymmetric graph to model the fusion similarity and then embeds it into the hash codes. However, matrix factorization-based methods can explore the correlation in multimodal data through the potential semantic space, which can avoid the high training complexity of calculating similarity graphs. Collective matrix factorization hashing (CMFH) [2] is a typical method based on matrix factorization, which learns the common representation from different modality data, and then quantizes it to obtain their hash codes. Collective reconstructive embedding (CRE) [17] employs different schema-specific modalities to handle heterogeneous data, which can dispose of the complex structural and heterogeneity of multi-modality data.

With the continuous development of deep learning, many deep hashing approaches have also been for unsupervised cross-modal retrieval. Liong et al. [18] proposed a three-layer neural network structure, which seeks multi-level nonlinear transformations to learn binary codes. Lin et al. [19] put forword to learn discriminative hash codes by introducing three criterion terms in the last layer of the network. Do et al. [20] designed a novel deep hashing network to efficiently learn hash codes by relaxing binary constraints. Similarity adaptive deep hashing (SADH) [21] alternately trains three modules: similarity graph updating, deep hashing model training and hash code optimization to obtain high-quality hash codes. Multi-pathway generative adversarial hashing (MGAH) [22] makes full use of the representation learning advantages of generative adversarial networks on unsupervised data to explore the latent manifold structure of cross-modal data. Deep graph-neighbor coherence preserving network (DGCPN) [23] was derived from the graph model to exploit the consistency of the neighbor graph by integrating the structure information between the data and its neighbors.

2.2 Supervised cross-modal hashing

Different from the aforementioned unsupervised hashing methods, supervised hashing methods try to fully exploit more semantic correlation from supervised information to improve retrieval accuracy. Cross-modality metric learning using similarity-sensitive hashing (CMSSH) [24] employs a binary classification approach to generate hash codes and employs an enhanced strategy to optimize the model. Supervised matrix factorization hashing (SMFH) [25] preserves the similarity by constructing an adjacency matrix and then employs relaxed discrete constraint to learn binary representation. Fast discrete cross-modal hashing (FDCH) [26] regresses category information to learn hash codes and hash functions. Liu et al; [27] proposed a universal and flexible cross-modal hashing framework, which can handle various cross-modal retrieval scenarios, including paired or unpaired multi-modal data retrieval and retrieval scenarios with equal or variable hash length coding. Different from the linear projection from Hamming space to label space, subspace relation in semantic labels for cross-modal hashing (SRLCH) [28] learns the linear transformation from label space to Hamming space by reverse learning. Its essence is to regard label information as advanced features and embeds it into hash codes.

Deep neural networks have also been widely used in supervised cross-modal retrieval due to their powerful arbitrary nonlinear representation capabilities. Deep cross-modal hashing (DCMH) [29] generates hash codes that preserve cross-modal similarity by imposing a negative log-likelihood loss in an end-to-end deep learning framework. Adversarial cross-modal retrieval (ACMR) [30] utilizes an adversarial learning classification approach to distinguish different modalities and generate binary hash codes. Cross-modal deep variational hashing (CMDVH) [31] put forward to a two-step framework to separate hash code learning and hash function generation. In the first step, CMDVH learns the unified hash codes of the image-text pairs in the database. Then it uses the learned unified hash codes to generate hash functions in the second step. Therefore, the learned hash function in the second stage cannot guide the optimization of the unified hash codes. Wang et al. [32] proposed a deep semantic reconstruction hash method with pairwise similarity-preserving quantitative constraints. This method embeds advanced semantic affinity in each data pair to learn compact binary codes.

3 Our proposed method

3.1 Notations

This proposed method adopts the batch strategy to train the model, where the variables are represented in a batch-wise manner. Specifically, let $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots., \boldsymbol{o}_k\}$ represent k instances in each batch, where $o_i = [\boldsymbol{I}_i, \boldsymbol{T}_i]$ is the i-th image-text pair. $X_I \in \mathbb{R}^{k\times d_1}$ and $X_T \in \mathbb{R}^{k\times d_2}$ denote the feature matrices of I_i and T_i, respectively. Generally, the image feature dimension d_1 and the text feature dimension d_2 should satisfy $d_1 \neq d_2$. Besides, let $\boldsymbol{B}_I \in \{-1, +1\}^{k\times l}$ and $\boldsymbol{B}_T \in \{-1, +1\}^{k\times l}$ represent the hash codes generated for image and text modality, respectively, where l is the hash code length. In addition, the label matrix is defined as $L \in R^{k\times c}$, where c represents the total category number.

3.2 Framework of our fine-grained similarity semantic preserving deep hashingmethod

3.2.1 Deep hashing networks for image and text modalities

The framework of the FSSPDH method is shown in Figure 1, which mainly consists of two parts: image hashing network and text hashing network. This network model can not only extract feature representations containing more semantic information for the two modalities of images and text, but also establish semantic

relationships between the two modalities through a semantic similarity matrix.

1) Image hashing network (ImgNet). As traditional SIFT features are not sufficient to capture the intrinsic semantic relationships of images, the proposed model follows previous work in extracting deep features from CNNs (pre-trained on ImageNet) to replace SIFT features. Thus, 4096-dimensional features are extracted from the fc7 layer (after ReLU) of AlexNet [33] as the image features for each input image block. Therefore, we use AlexNet as the backbone of ImgNet and replace the classifier layer fc8 of AlexNet with a new fc with l hidden units to generate a continuous representation $H_I \in \mathbb{R}^{k\times l}$.

2) Text hashing network (TxtNet). For the text modality, LDA (Latent Dirichlet Allocation) topic vectors or token features are as X_T. In addition, we use multilayer perceptron (MLP) as the backbone of TxtNet. Due to the diversity and complexity of raw text descriptions, we directly use topic vectors or label occurrence features as the input of the MLP and then have 4096 units in the first fc layer. Besides, the second fc layer with l units generates a continuous representation $H_T \in \mathbb{R}^{k\times l}$ and ReLU is used as the activation function.

3.2.2 Constructing fine-grained similarity matrix

To improve the retrieval performance of cross-modal hashing methods using supervised information, most methods usually adopt the labels to construct a high-level semantic similarity matrix. Specifically, the high-level semantic similarity $S_H \in \{-1,+1\}^{k\times k}$ is computed by $\boldsymbol{S}_H = \boldsymbol{L}\boldsymbol{L}^T$. If the i-th and j-th samples share at least one label, then $\boldsymbol{S}_{H_{ij}} = 1$, otherwise $\boldsymbol{S}_{H_{ij}} = -1$. For multi-label datasets, samples with multiple labels should be more similar than these with only one label. However, the similarity only based on labels cannot effectively model this relationship, and a lot of useful information is discarded. To solve this issue, we construct a high-level similarity S_H and a low-level similarity S_L using labels and features, respectively. Therefore, samples with the same high-level similarity can be further ranked according to their low-level similarity. The construction of the fine-grained similarity can be expressed as:

$$S = \mu S_H + \theta S_L, \tag{1}$$

where μ and θ are used to balance high-level similarity and low-level similarity. In addition, according to the fine-grained similarity fusion rules described in Ref. [34], the fine-grained similarity matrix can be represented as follows:

$$S = \frac{\mu\boldsymbol{L}\boldsymbol{L}^T + \theta_1\boldsymbol{X_t}\boldsymbol{X_I}^T + \theta_2\boldsymbol{X_T}\boldsymbol{X_T}^T}{\mu + 1}, \tag{2}$$

where θ_1 and θ_2 are the weight parameters of image and text, respectively.

3.2.3 Hash codes learning

The goal of our FSSPDH method is to learn different hash codes for different modalities and establish relationships between modalities and instances by similarity matrix. FSSPDH seeks to map the features of instances to the Hamming space that preserves semantic similarity. In this space, the hash codes of samples from the same category should be similar. However, the hash codes of samples from different categories should also be different. Therefore, we attempt to preserve the semantic similarity between the hash codes learned from different modalities and the hash codes learned from the same instance of the same modality. Specifically, if $\boldsymbol{S}_{ij} = 1$ indicates that the hash codes b_i and b_j are similar, the Hamming distance between b_i and b_j should be the minimum value of 0, which means that $b_i^T b_j = c$. Otherwise, the Hamming distance between b_i and b_j should be the minimum value of c, which means that $b_i^T b_j = 0$. In the training stage, to calculate the gradient in backpropagation, we use the scaled tanh function to obtain approximate hash codes [35]. Therefore, B_I and B_T in the training phase can be calculated by the following formulas:

$$\boldsymbol{B}_I = \tanh(\alpha\boldsymbol{H}_I) \in [-1,+1]^{k\times l}, \tag{3}$$

$$\boldsymbol{B}_T = \tanh(\alpha\boldsymbol{H}_T) \in [-1,+1]^{k\times l}, \tag{4}$$

where α is a smooth parameter and needs to satisfy the following constraint: $\lim_{\alpha\to 0}\tanh(\alpha x) = sgn(x)$. $sgn(\cdot)$ is a symbolic function.

1) Fine-grained similarity semantic preserving learning. Our FDSSPH method considers both the inter-modality similarity and intra-modality similarity to guide the learning of hash codes. Therefore, we use Mean Square Error (MSE) to define the hash loss:

TABLE 1 The MAP values of cross-modal retrieval on WIKI dataset.

Task	Methods	WIKI			
		16	32	64	128
T2I	CVH	-	-	-	-
	JIMFH	0.4024	0.4564	0.4630	0.4695
	DCH	0.6366	0.6417	0.6518	0.6500
	DLFH	0.4268	0.5836	0.6109	0.6478
	DCMH	0.5553	0.5742	0.5984	0.5876
	SSAH	-	-	-	-
	DCHUC	0.5224	0.5047	0.5561	0.6392
	DJRSH	0.3337	0.3633	0.3782	0.3981
	FSSPDH	**0.6528**	**0.6850**	**0.6614**	**0.6650**
I2T	CVH	-	-	-	-
	JIMFH	0.1430	0.1272	0.1314	0.1353
	DCH	0.2115	0.2298	0.2354	0.2443
	DLFH	0.1858	0.2090	0.2269	0.2312
	DCMH	0.3655	0.3792	0.3842	0.3794
	SSAH	-	-	-	-
	DCHUC	0.2358	0.2490	0.2822	0.3066
	DJRSH	0.2756	0.2788	0.3043	0.3148
	FSSPDH	**0.3753**	**0.4044**	**0.3935**	**0.4054**

"–" denotes an untested value under that specific setting. The bold value mean the best performance.

$$\Gamma_s = \|\boldsymbol{S} - \cos(\boldsymbol{B}_I, \boldsymbol{B}_T)\|_F^2 + \beta_1\|\boldsymbol{S} - \cos(\boldsymbol{B}_I, \boldsymbol{B}_I)\|_F^2 + \beta_2\|\boldsymbol{S} - \cos(\boldsymbol{B}_T, \boldsymbol{B}_T)\|_F^2 \quad s.t. \quad \boldsymbol{B}_I, \boldsymbol{B}_T \in [-1, +1]^{k\times l}, \tag{5}$$

where β_1 and β_2 are the balance parameters of intra-modality similarity learning items.

2) Quantized loss learning. Hash loss defined by MSE can generate modal-specific hash representations B_I and B_T. However, there are differences between hash codes and hash representations. Therefore, we add a quantization loss to reduce the information loss from hash representations to hash codes. The quantization loss function can be defined as follows:

$$\Gamma_q = \lambda\left(\left\|sgn(\boldsymbol{B}_I) - \boldsymbol{B}_I\right\|_F^2 + \left\|sgn(\boldsymbol{B}_T) - \boldsymbol{B}_T\right\|_F^2\right) \quad s.t.\ \boldsymbol{B}_I, \boldsymbol{B}_T \in [-1, +1]^{k\times l}, \tag{6}$$

where λ is a non-negative parameter, and its role is to balance the weight of the quantization loss term.

3.2.4 Overall objective function

By integrating Eqs 5, 6 into a unified framework, the overall objective function of the proposed FSSPDH approach is given as follows:

$$\begin{aligned}\min_{\boldsymbol{B}_I, \boldsymbol{B}_T} \Gamma &= \Gamma_s + \Gamma_q \\ &= \|\boldsymbol{S} - \cos(\boldsymbol{B}_I, \boldsymbol{B}_T)\|_F^2 + \beta_1\|\boldsymbol{S} - \cos(\boldsymbol{B}_I, \boldsymbol{B}_I)\|_F^2 + \beta_2\|\boldsymbol{S} - \cos(\boldsymbol{B}_T, \boldsymbol{B}_T)\|_F^2 \\ &+ \lambda\left(\left\|sgn(\boldsymbol{B}_I) - \boldsymbol{B}_I\right\|_F^2 + \left\|sgn(\boldsymbol{B}_T) - \boldsymbol{B}_T\right\|_F^2\right) \\ &s.t.\ \boldsymbol{B}_I, \boldsymbol{B}_I \in [-1, +1]^{k\times l}.\end{aligned} \tag{7}$$

Algorithm 1 describes the overall training process of our proposed FSSPDH approach in detail.

Input: The feature matrices X_I and X_T, the label matrix L of the trainingset $\{\boldsymbol{o}_i = [\boldsymbol{I}_i, \boldsymbol{T}_i]\}_{i=1}^n$, the hash code length l and the parameters ψ_{θ_1}, ψ_{θ_2} of TxtNetnetwork and ImgNet network, the size k of training batch.
Output: Hash functions of image and text modalities.
Procedure:
1. Initialize $t = 0$.
Repeat
2 $t = t + 1, \alpha = \sqrt{t}$
3 **For** all training samples enter the model **do**
4 Randomly select k samples from the training set.
5 Calculate the fine-grained similarity matrix S by Eq. 2.
6 Forward propagation $H_I = \psi_{\theta_1}(X_I)$ and $H_T = \psi_{\theta_2}(X_T)$.
7 Calculate hash representations B_I and B_T of image and text modalities by Eqs 3, 4.
8 Calculate overall objective function Γ by Eq. 7.
9 Update the whole parameters using back-propagating gradient by chain rule.
10 **End for**
Until convergence.

Algorithm 1 FSSPDH.

3.3 Out-of-sample problem

Since our proposed method can only obtain the hash codes of training data, it cannot effectively solve the out-of-samples problem. Therefore, it is still necessary to generate the hash codes of the query samples that are absent in the training set. To solve this problem, we can obtain the hash codes of query sample x_q by forward propagation

$$b_q = sgn\left(\tanh\left(\psi_\theta\left(\boldsymbol{x}_q\right)\right)\right). \tag{8}$$

4 Experiments

4.1 Datasets

The **WIKI** [36] dataset consists of 2,866 image-text pairs belonging to 10 different categories. For this experiment, the entire dataset was used as the retrieval dataset, with 2,173 pairs used for training and the remaining 693 pairs used for querying.

The **MIRFLICKR-25K** [37] dataset is a multi-label dataset obtained from the FLICKR website. In this experiment, 20015 samples were selected as experimental samples, each of which is tagged with at least one of the 24 categories. In this experiments, 2,243 samples were randomly selected as query samples, and the remaining 17772 samples were used as retrieval samples. From the retrieval samples, 5,000 samples were selected for training.

The **NUSWIDE** [38] dataset is a multimodal dataset consisting of 269648 image-text pairs, each of which corresponds to at least one or more of the 81 categories. Here, the most common 21 categories and their corresponding 195749 samples were selected to evaluate the effectiveness of the proposed FSSPDH approach. From these experimental data, 2000 samples were randomly selected as query samples, and the remaining samples were used as retrieval samples. Besides, 10000 samples were selected from the retrieval samples for traning model.

4.2 Baselines and implementation details

To demonstrate the superiority of the FDSSPH method, we compared it with several mainstream hashing methods, such as cross-view hashing (CVH) [39], joint and individual matrix factorization hashing (JIMFH) [40], discrete cross-modal hashing (DCH) [41], discrete latent factor model for cross-modal hashing (DLFH) [3], DCMH [29], SSAH [13], DCHUC [12], and deep joint-semantics reconstructing hashing (DJRSH) [42]. Besides, we evaluated these hashing methods on the WIKI, MIRFLICKR-25K and NUSWIDE datasets for both image-to-text (I2T) and text-to-image (T2I) retrieval tasks. The lengths of hash codes were set to 16, 32, 64, and 128 bits, respectively. The hyperparameters in the model were set to $\beta_1 = 0.1$, $\beta_2 = 0.1$ and $\lambda = 0.01$ according to our empirical knowledge.

4.3 Evaluation

In this paper, mean average precision (MAP) and TopN-precision curves are used to evaluate the performances of the proposed method and baseline methods. MAP is one of the most metrics in cross-modal

TABLE 2 The MAP values of cross-modal retrieval on MIRFLICKR-25K dataset.

Task	Methods	MIRFLICKR-25K			
		16	32	64	128
T2I	CVH	0.6240	0.6323	0.6364	0.6374
	JIMFH	0.6659	0.6591	0.6424	0.6900
	DCH	0.7246	0.7546	0.7730	0.8028
	DLFH	0.7795	0.8059	0.8262	0.8379
	DCMH	0.7993	0.8117	0.8218	0.8206
	SSAH	0.8286	0.8311	0.8338	0.8251
	DCHUC	0.7745	0.7939	0.8202	0.8207
	DJRSH	0.6317	0.7213	0.7590	0.7733
	FSSPDH	**0.8558**	**0.8509**	**0.8559**	**0.8653**
I2T	CVH	0.6174	0.6154	0.6154	0.6129
	JIMFH	0.6506	0.6453	0.3657	0.6862
	DCH	0.6647	0.6865	0.7063	0.7268
	DLFH	0.6803	0.7002	0.7158	0.7310
	DCMH	0.7704	0.7581	0.8073	0.8104
	SSAH	0.8236	0.8296	0.8450	0.8662
	DCHUC	0.7619	0.7953	0.8162	0.8176
	DJRSH	0.7133	0.7605	0.7889	0.7979
	FSSPDH	**0.8268**	**0.8496**	**0.8691**	**0.8776**

The bold value mean the best performance.

TABLE 3 The MAP values of cross-modal retrieval on NUSWIDE dataset.

Task	Methods	NUSWIDE			
		16	32	64	128
T2I	CVH	0.5820	0.5734	0.5621	0.536
	JIMFH	0.6337	0.6704	0.6916	0.7123
	DCH	0.7028	0.7205	0.7687	**0.7839**
	DLFH	0.6662	0.7445	0.7569	0.7686
	DCMH	0.6845	0.6931	0.7053	0.7067
	SSAH	0.6734	0.6621	0.6206	0.6445
	DCHUC	0.6491	0.6973	0.7178	0.6982
	DJRSH	0.5629	0.7019	0.7027	0.7694
	FSSPDH	**0.7154**	**0.7712**	**0.7816**	0.7766
I2T	CVH	0.5561	0.5452	0.5383	0.5201
	JIMFH	0.6528	0.6719	0.6802	0.6875
	DCH	0.6174	0.6752	0.6849	0.6854
	DLFH	0.6174	0.6752	0.6849	0.6854
	DCMH	0.6740	0.6901	0.7314	0.7611
	SSAH	0.6841	0.7054	0.7361	0.7334
	DCHUC	0.7469	0.7549	0.7911	0.7637
	DJRSH	0.6193	0.7173	0.7178	0.7936
	FSSPDH	**0.7554**	**0.7723**	**0.8059**	**0.7943**

The bold value mean the best performance.

retrieval tasks. Specially, the average precision of a given query sample and the returned results can be defined as follows:

$$AP = \frac{1}{n}\sum_{r=1}^{R} P(r)\delta(r), \tag{9}$$

where n is the number of true samples returned. $P(r)$ is the precision of the last r sample returned. If the returned sample is similar to the query sample, then $\delta(r) = 1$, otherwise $\delta(r) = 0$. In this experiment, R was empirically set to 1,000. In other words, the accuracy of the first 1,000 retrieved samples was reported. The MAP value is the average value of AP for all query samples, which is defined as follows:

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP\left(x_q\right), \tag{10}$$

where Q is the number of query samples. In addition, TopN-precision is defined as similar for the first N instances retrieved from all queries within the Hamming distance. In order to return the results of a more accurate group, we set N to 1,000.

4.4 Experimental results and discussion

In this section, we conducted different retrieval experiments on the three datasets to evaluate the proposed FSSPDH method and its competitors. Table 1, Table 2, Table 3 shows the MAP values of the proposed method and baseline methods on different multimedia datasets.

Table 1, Table 2, Table 3 show the mAP values of all methods on three datasets. Figure 2 shows the Top-N precision curves of the proposed approach and its competitors. Based on these retrieval results, we can draw some conclusions as follows.

1) It is clear from Table 1 that our proposed FSSPDH method outperforms other baseline methods on three multimedia datasets. Specifically, compared with the results with 128 bits, our FSSPDH method performs almost 2.2% better than the second best DLFH method in the T2I task on the WIKI dataset. On the MIRFLICKR dataset, our FSSPDH method performs nearly 2.0% better than the second best DLFH method. On the NUSWIDE dataset, the FSSPDH method has a performance improvement of almost 1.0% over the second-best method. Therefore, we can known from the retrieval results that the proposed FSSPDH method has greater advantages over other hashing methods in cross-modal retrieval tasks.
2) In addition, the experiments on the three different datasets also show that the FSSPDH approach improves the retrieval accuracy to some extent in the I2T task. Compared with the three data sets, we can find that the method on the MRIFLICKR-25K data set is higher than other two data sets. This is because the data

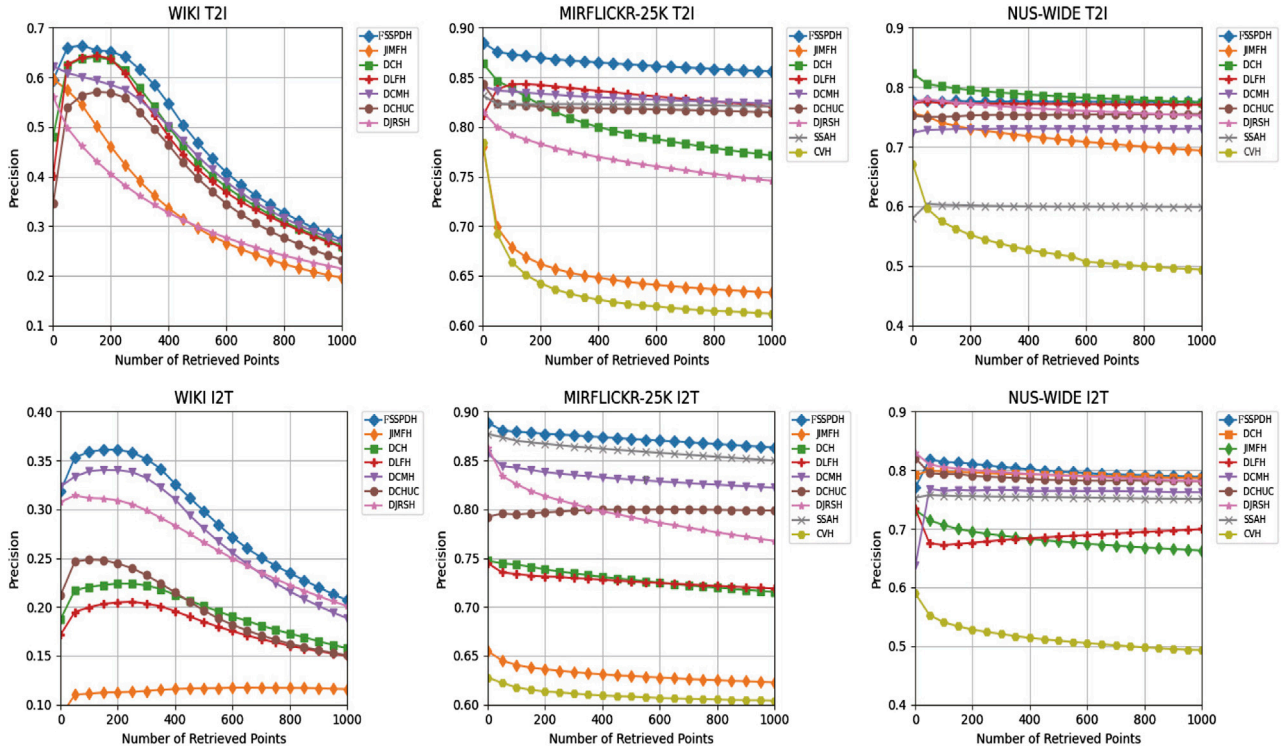


FIGURE 2
The TopN-precision curves of FSSPDH and its competitors on three datasets.

TABLE 4 Ablation results of our FSSPDH approach on the MIRFlickr dataset.

Methods	I2T	T2I
FSSPDH-II	0.8706	0.8736
FSSPDH-TT	0.8711	0.8690
FSSPDH-IT	0.5052	0.5649
FSSPDH-Q	0.8717	0.8705
FSSPDH	**0.8726**	**0.8746**

The bold value mean the best performance.

distribution, division and size of the data set can affect the retrieval performance of the proposed method.

3) We can see that all methods cannot achieve excellent performances on the WIKI dataset. This is because this dataset contains fewer samples and lower data dimensionality for text modality features. Therefore, most deep hashing approaches cannot fully leverage the advantages of deep learning and thus lead to poor retrieval performance in general. However, our proposed FSSPDH method can still achieve the best performance among all cross-modal retrieval methods on this dataset.
4) It can be found from the experimental results that the performances of most methods can improve with the increase of hash code length. The main reason is that the longer hash codes usually contains more semantic information. However, the performance of some methods decreases when the hash code length ranges from 64 bits to 128 bits. The possible reason is that the learned hash codes contains more useless information, which leads to the decline of retrieval performance.
5) It is clear to see from Figure 2 that our FSSPDH method achieves the best performance among the compared methods from the perspective of TopN-precision. In addition, we can observe that the TopN-precision curve results are basically consistent with the MAP value results, as they are both calculated based on the Hamming distance. This indicates that our proposed FSSPDH method also achieves the best results in the Hamming ranking task.
6) We can find that the TopN-precision curves of all methods on three datasets show slightly different decreasing rates. Specifically, the WIKI dataset includes the least amount of data, and its curve decline rate is obviously higher than these of the other two datasets. Note that the NUSWIDE dataset contains the most data, so its TopN-precision curve is relatively flat. However, our proposed method considers the semantic similarity between and within modalities by constructing a fine-grained similarity matrix, thereby achieving the best results on three different scale datasets.

4.5 Ablation experiment and analysis

To verify the effectiveness of each component in the proposed FSSPDH approach, we constructed four variants of FSSPDH, i.e., FSSPDH-II, FSSPDH-TT, FSSPDH-IT, and FSSPDH-Q.

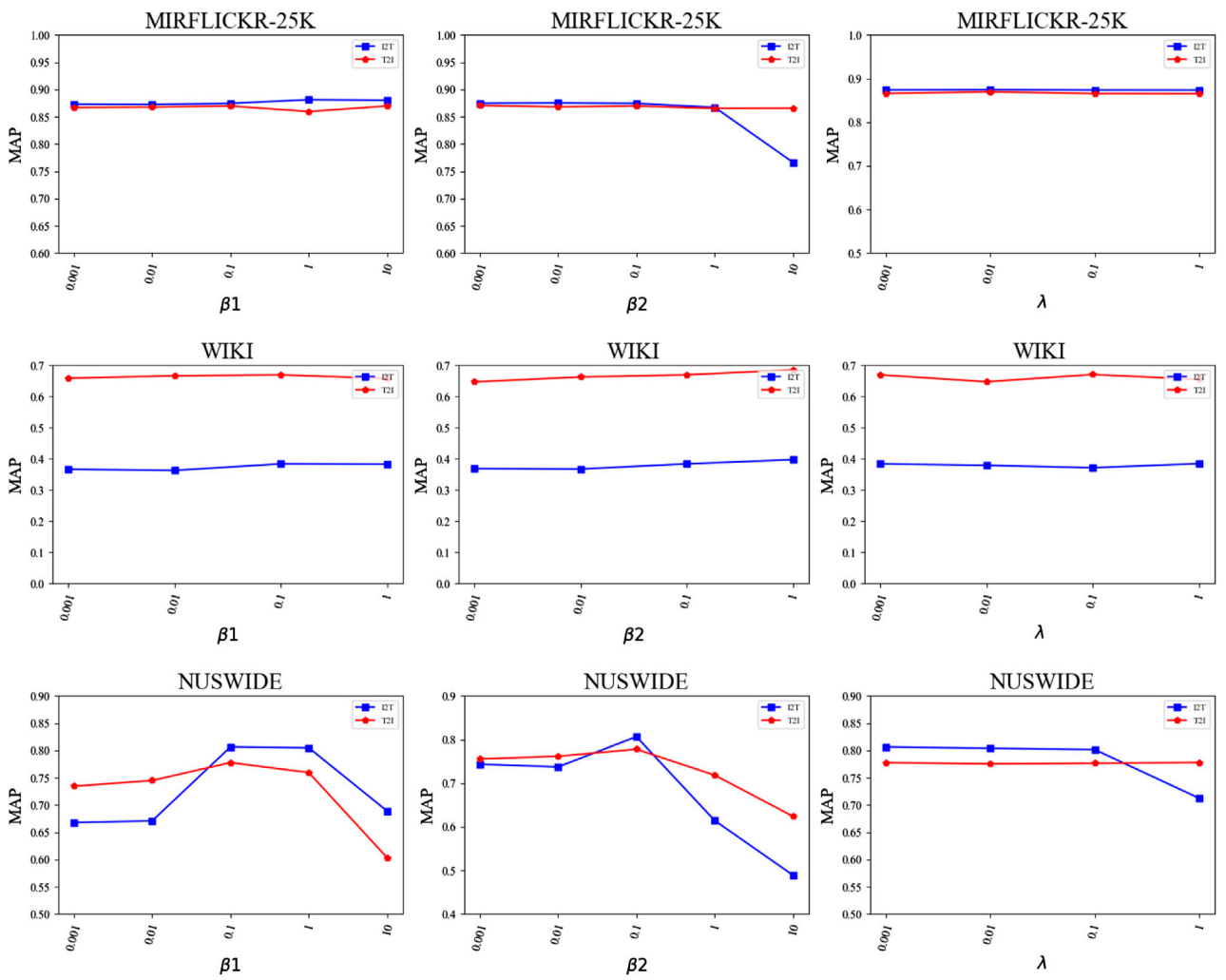


FIGURE 3
The retrieval performances of the FSSPDH method with different parameter settings.

FSSPDH-II was constructed by removing the intra-modal similarity learning for the image modality. FSSPDH-TT discarded the intra-modal similarity learning for the text modality. FSSPDH-IT removed the inter-modal similarity learning for both image and text modalities, while FSSPDH-Q discarded the hashing quantization loss term. These ablation experiments were conducted on the MIRFlickr dataset to validate the impact of each component on retrieval performance. Here, the hash code length was set to 128 bits in this experiments. Table 4 shows the retrieval performances of FSSPDH and its variants on two retrieval tasks.

It can be seen from Table 4 that FSSPDH-IT cannot outperform other variants on different retrieval tasks, which indicates that inter-modal similarity learning is crucial for retrieval performance in our method. In addition, the performances of the FSSPDH-II, FSSPDH-TT, and FSSPDH-Q variants are also lower than that of FSSPDH in different retrieval tasks. It shows that both intra-modal similarity learning and hashing quantization loss can be beneficial in enhancing retrieval performance.

4.6 Parameter sensitivity analysis

Our FSSPDH method mainly includes three hyperparameters: β_1, β_2 and λ. This subsection discusses the impact of different hyperparameter values in our proposed model. In this experiment, the length of the hash codes was designated as 128 bits. Specifically, we change the values of only one hyperparameter by fixing the values of the other two hyperparameters. Figure 3 plots the results of the proposed FSSPDH approach with different parameter settings on three datasets. We can see from Figure 3 that the performances of FSSPDH on the WIKI dataset and MIRFLICKR-25K dataset are relatively stable within a large range of hyperparameter values. Besides, our FSSPDH approach has fluctuated to a certain extent on the NUSWIDE dataset with different hyperparameter values. Fortunately, we can see that the FSSPDH approach can also obtain relatively stable performances within a certain range. Therefore, it can be found that our FSSPDH approach is insensitive to the hyperparameters from the parameter experiments.

5 Conclusion

In this paper, we introduce a novel approach called fine-grained similarity semantic preserving deep hashing (FSSPDH) for cross-modal retrieval. Firstly, the FSSPDH approach attempts to learn a set of binary hash codes for each modality and thus effectively preserves the characteristics of each modality. In addition, our FSSPDH approach constructs a fine-grained semantic similarity matrix by using labels and features, which not only preserves the inter-modal similarity but also maintains the intra-modal similarity. Therefore, the fine-grained similarity preserving strategy is used to embed more semantic information into hash codes. Compared with other hashing methods, it can preserve the inter-modality similarity and maintain the semantic relationships between instances by the intra-modality similarity, simultaneously, thus narrowing the heterogeneous gap between different modalities. Additionally, to reduce the information loss from the continuous hash representation to discrete hash codes, our FSSPDH approach incorporates hash quantization loss to further improve the retrieval performance. A series of experimental results have demonstrated that the proposed FSSPDH method achieves superior performances in cross-modal retrieval tasks on different multimedia datasets.

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

Conceptualization, GL; methodology, QP; validation, GL, QP, DZ, and ZS; formal analysis, JY; writing—review and editing, ZS; investigation, ZS; resources, ZS All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by National Natural Science Foundation of China (61603159, 62162033) and Yunnan Foundation Research Projects (202201AT070154, 202101BE070001-056).

Conflict of interest

Author GL was employed by Yunnan Power Grid Corporation, China. Authors QP, DZ, and JY were employed by Electric Power Research Institute, Yunnan Power Grid Corporation, China.

The remaining author declares 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. Kaur P, Pannu HS, Malhi AK. Comparative analysis on cross-modal information retrieval: A review. *Comp Sci Rev* (2021) 39:100336. doi:10.1016/j.cosrev.2020.100336

2. Ding G, Guo Y, Zhou J. Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2014). p. 2075–82.

3. Jiang Q-Y, Li W-J. Discrete latent factor model for cross-modal hashing. *IEEE Trans Image Process* (2019) 28:3490–501. doi:10.1109/tip.2019.2897944

4. Shu Z, Yong K, Yu J, Gao S, Mao C, Yu Z. Discrete asymmetric zero-shot hashing with application to cross-modal retrieval. *Neurocomputing* (2022) 511:366–79. doi:10.1016/j.neucom.2022.09.037

5. Song J, Yang Y, Yang Y, Huang Z, Shen HT. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the 2013 ACM SIGMOD international conference on management of data (2013). p. 785–96.

6. Zhou J, Ding G, Guo Y. Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (2014). p. 415–24.

7. Shu Z, Li L, Yu J, Zhang D, Yu Z, Wu X-J. Online supervised collective matrix factorization hashing for cross-modal retrieval. *Appl intelligence* (2022) 1–18. doi:10.1007/s10489-022-04189-6

8. Shu Z, Yong K, Zhang D, Yu J, Yu Z, Wu X-J. Robust supervised matrix factorization hashing with application to cross-modal retrieval. *Neural Comput Appl* (2023) 35:6665–84. doi:10.1007/s00521-022-08006-6

9. Deng C, Chen Z, Liu X, Gao X, Tao D. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans Image Process* (2018) 27:3893–903. doi:10.1109/tip.2018.2821921

10. Wang X, Zou X, Bakker EM, Wu S. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval. *Neurocomputing* (2020) 400:255–71. doi:10.1016/j.neucom.2020.03.019

11. Shu Z, Bai Y, Zhang D, Yu J, Yu Z, Wu X-J. Specific class center guided deep hashing for cross-modal retrieval. *Inf Sci* (2022) 609:304–18. doi:10.1016/j.ins.2022.07.095

12. Tu R-C, Mao X-L, Ma B, Hu Y, Yan T, Wei W, et al. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Trans Knowledge Data Eng* (2020) 34:560–72. doi:10.1109/tkde.2020.2987312

13. Li C, Deng C, Li N, Liu W, Gao X, Tao D. Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018). p. 4242–51.

14. Zhu X, Huang Z, Shen HT, Zhao X. Linear cross-modal hashing for efficient multimedia search. In: Proceedings of the 21st ACM international conference on Multimedia (2013). p. 143–52.

15. Zheng F, Tang Y, Shao L. Hetero-manifold regularisation for cross-modal hashing. *IEEE Trans pattern Anal machine intelligence* (2016) 40:1059–71. doi:10.1109/tpami.2016.2645565

16. Liu H, Ji R, Wu Y, Huang F, Zhang B. Cross-modality binary code learning via fusion similarity hashing. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017). p. 7380–8.

17. Hu M, Yang Y, Shen F, Xie N, Hong R, Shen HT. Collective reconstructive embeddings for cross-modal hashing. *IEEE Trans Image Process* (2018) 28:2770–84. doi:10.1109/tip.2018.2890144

18. Erin Liong V, Lu J, Wang G, Moulin P, Zhou J. Deep hashing for compact binary codes learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2015). p. 2475–83.

19. Lin K, Lu J, Chen C-S, Zhou J. Learning compact binary descriptors with unsupervised deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016). p. 1183–92.

20. Do T-T, Doan A-D, Cheung N-M. Learning to hash with binary deep neural network. In: Computer Vision–ECCV 2016: 14th European Conference; October 11-14, 2016; Amsterdam, The Netherlands. Springer (2016). p. 219–34.

21. Shen F, Xu Y, Liu L, Yang Y, Huang Z, Shen HT. Unsupervised deep hashing with similarity-adaptive and discrete optimization. *IEEE Trans Pattern Anal Machine Intelligence* (2018) 40:3034–44. doi:10.1109/tpami.2018.2789887

22. Zhang J, Peng Y. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Trans Multimedia* (2019) 22:174–87. doi:10.1109/tmm.2019.2922128

23. Yu J, Zhou H, Zhan Y, Tao D. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. *Proc AAAI Conf Artif Intelligence* (2021) 35: 4626–34. doi:10.1609/aaai.v35i5.16592

24. Bronstein MM, Bronstein AM, Michel F, Paragios N. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: 2010 IEEE computer society conference on computer vision and pattern recognition; 05 August 2010; San Francisco, CA, USA. IEEE (2010). p. 3594–601.

25. Tang J, Wang K, Shao L. Supervised matrix factorization hashing for cross-modal retrieval. *IEEE Trans Image Process* (2016) 25:3157–66. doi:10.1109/tip.2016.2564638

26. Liu X, Nie X, Zeng W, Cui C, Zhu L, Yin Y. Fast discrete cross-modal hashing with regressing from semantic labels. In: Proceedings of the 26th ACM international conference on Multimedia (2018). p. 1662–9.

27. Liu X, Hu Z, Ling H, Cheung Y-m. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE Trans Pattern Anal Machine Intelligence* (2019) 43:964–81. doi:10.1109/tpami.2019.2940446

28. Shen HT, Liu L, Yang Y, Xu X, Huang Z, Shen F, et al. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Trans Knowledge Data Eng* (2020) 33: 3351–65. doi:10.1109/tkde.2020.2970050

29. Jiang Q-Y, Li W-J. Deep cross-modal hashing. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017). p. 3232–40.

30. Wang B, Yang Y, Xu X, Hanjalic A, Shen HT. Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM international conference on Multimedia (2017). p. 154–62.

31. Erin Liong V, Lu J, Tan Y-P, Zhou J. Cross-modal deep variational hashing. In: Proceedings of the IEEE international conference on computer vision (2017). p. 4077–85.

32. Wang Y, Ou X, Liang J, Sun Z. Deep semantic reconstruction hashing for similarity retrieval. *IEEE Trans Circuits Syst Video Tech* (2020) 31:387–400. doi:10.1109/tcsvt.2020.2974768

33. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM* (2017) 60:84–90. doi:10.1145/3065386

34. Wang Y, Chen Z-D, Luo X, Xu X-S. A high-dimensional sparse hashing framework for cross-modal retrieval. *IEEE Trans Circuits Syst Video Tech* (2022) 32: 8822–36. doi:10.1109/tcsvt.2022.3195874

35. Li X, Hu D, Nie F. Deep binary reconstruction for cross-modal hashing. In: Proceedings of the 25th ACM international conference on Multimedia (2017). p. 1398–406.

36. Pereira JC, Coviello E, Doyle G, Rasiwasia N, Lanckriet GR, Levy R, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Trans pattern Anal machine intelligence* (2013) 36:521–35.

37. Huiskes MJ, Lew MS. The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval (2008). p. 39–43.

38. Chua T-S, Tang J, Hong R, Li H, Luo Z, Zheng Y. Nus-wide: A real-world web image database from national University of Singapore. In: Proceedings of the ACM international conference on image and video retrieval (2009). p. 1–9.

39. Kumar S, Udupa R. Learning hash functions for cross-view similarity search. In: Twenty-second international joint conference on artificial intelligence (2011).

40. Wang D, Wang Q, He L, Gao X, Tian Y. Joint and individual matrix factorization hashing for large-scale cross-modal retrieval. *Pattern recognition* (2020) 107:107479. doi:10.1016/j.patcog.2020.107479

41. Xu X, Shen F, Yang Y, Shen HT, Li X. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Trans Image Process* (2017) 26:2494–507. doi:10.1109/tip.2017.2676345

42. Su S, Zhong Z, Zhang C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision (2019). p. 3027–35.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 10 May 2023
DOI 10.3389/fphy.2023.1187503

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Yunchun Zhang,
Yunnan University, China
Zhongqing Wang,
Soochow University, China

*CORRESPONDENCE
Junjun Guo,
guojjgb@163.com

RECEIVED 16 March 2023
ACCEPTED 12 April 2023
PUBLISHED 10 May 2023

CITATION
Xiang Y, Cai Y and Guo J (2023), MSFNet: modality smoothing fusion network for multimodal aspect-based sentiment analysis.
Front. Phys. 11:1187503.
doi: 10.3389/fphy.2023.1187503

MSFNet: modality smoothing fusion network for multimodal aspect-based sentiment analysis

Yan Xiang[1,2], Yunjia Cai[1,2] and Junjun Guo[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

Multimodal aspect-based sentiment classification (MABSC) aims to determine the sentiment polarity of a given aspect in a sentence by combining text and image information. Although the text and the corresponding image in a sample are associated with aspect information, their features are represented in distinct semantic spaces, creating a substantial semantic gap. Previous research focused primarily on identifying and fusing aspect-level sentiment expressions of different modalities while ignoring their semantic gap. To this end, we propose a novel aspect-based sentiment analysis model named modality smoothing fusion network (MSFNet). In this model, we process the unimodal aspect-aware features via the feature smoothing strategy to partially bridge modality gap. Then we fuse the smoothed features deeply using the multi-channel attention mechanism, to obtain aspect-level sentiment representation with comprehensive representing capability, thereby improving the performance of sentiment classification. Experiments on two benchmark datasets, Twitter2015 and Twitter2017, demonstrate that our model outperforms the second-best model by 1.96% and 0.19% in terms of Macro-F1, respectively. Additionally, ablation studies provide evidence supporting the efficacy of each of our proposed modules. We release the code at: https://github.com/YunjiaCai/MSFNet.

KEYWORDS

multimodal sentiment analysis, aspect-based sentiment analysis, multimodal fusion, feature smoothing, semantic gap

1 Introduction

In recent years, there has been a significant increase in the amount of multimodal data from various social, shopping, and news platforms. These data consist primarily of a piece of text and an associated image, and are often accompanied by a personal sentiment tendency. Analyzing the sentiment towards specific aspects in this type of data can provide valuable insights into people's personalized preferences or predict public opinion trends. Therefore, multimodal aspect-based sentiment classification (MABSC) has received extensive attention. The objective of this task is to combine a piece of text, its associated image and a given aspect from the text to determine the sentiment polarity of the given aspect. As shown in Figure 1, the sentiment polarity of the aspect {*Rocky*} could be determined as {*Neutral*}, according to the text alone. However, by combining image information, it can be determined that the aspect term has a {*Positive*} sentiment polarity. Therefore, the key to this task lies in effectively extracting and combining the sentiment features from both images and texts.

From the feature learning perspective, images and texts are commonly represented in distinct feature spaces, which creating a semantic gap between the two modalities and posing substantial

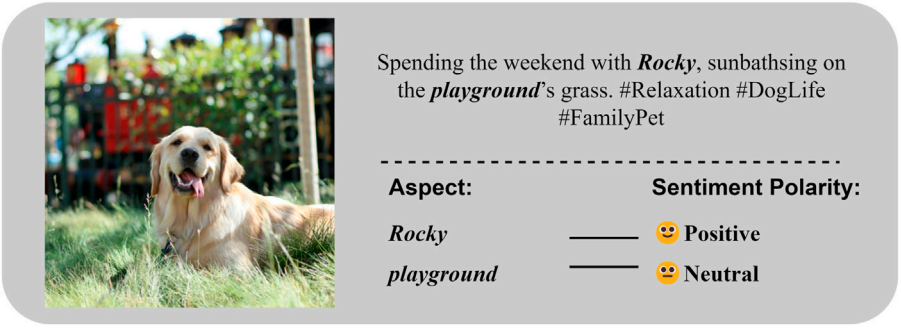


FIGURE 1
Example of MABSC tasks.

challenges for subsequent inter-modal interactions [1, 2]. As a result, the major difficulty of MABSC is to bridge the gap between modalities and model the deep interactions of them. Early MABSC research primarily relied on directly modeling the interaction between modalities to achieve multimodal fusion. Xu et al. [3]proposed a memory-based model which extracted text and image features using pre-trained Bert and ResNet models respectively, and stacked interactive attention mechanism with several memory hops to learn the deep abstraction of multimodal data. Similarly, Zhang et al. [4] sent features of two different modalities into a fusion discriminant matrix to learn the interaction of different modalities and a similarity matrix is used to capture modal invariant features, based on which the consistency and redundancy of different modalities can be identified. However, the deficiency of these methods was that they did not consider the semantic gaps on subsequent interactions. Khan et al. [1] recognized the influence of semantic gaps on multimodal fusion and used a cross-modal Transformer to map image content to the text space. They then utilized a pre-trained Bert structure to model the interactions between image, text, and aspect. However, the performance is limited due to the lack of in-depth exploration of inter-modal interactions.

To tackle the problem of insufficient fusion, we propose a novel MABSC model called "modality smoothing fusion network (MSFNet)". The main contribution can be summarized as follows.

- Unlike existing works of MABSC that mainly study extracting and fusing aspect-level sentiment expressions, we focus on the problem that modality discrepancy influnce their subsequent fusions.
- The proposed MSFNet adopts the feature smoothing strategy and the multi-channel attention to effectively bridge the semantic gap and achieve better fusion of text-image modalities.
- Experimental results on two benchmark datasets verify that MSFNet achieve effective interaction of multimodalities and obtains state-of-the-art performance in MABSC.

2 Related work

Aspect-based sentiment classification (ABSC) was first proposed on text datasets. With the increase of multimodal data, multimodal sentiment analysis (MSA) gained great attention, and MABSC is the research combining ABSC and MSA.

2.1 Aspect-based sentiment classification (ABSC)

Aspect-Based Sentiment Classification (ABSC) is a task that involved predicting the sentiment polarity of a target entity within a given sentence. Traditional methods for ABSC relied on manually annotated features, such as language rules [5] and feature engineering [6]. In recent years, neural networks have shown great promise in this area and have led to significant performance improvements. Early neural network approaches typically used Long Short-Term Memory to model the interaction between the aspect and its context [7]. More recent works have incorporated attention mechanisms to select aspect-related sentiment features [8], with some studies introducing more complex interactive attention methods to learn aspect-specific representations [9, 10]. These methods demonstrate the significance of contextual information in the task of aspect sentiment analysis. Pre-trained language models, such as BERT [11], have also been utilized to improve the ABSC performance [12].

2.2 Multimodal sentiment Analysis (MSA)

MSA aims to combine multimodal information such as text, visual, and audio to understand human emotions [13]. Previous researchers have primarily focused on unimodal representation learning and multimodal fusion.

Unimodal representation learning: Wang et al. [14] constructed a recurrent variational embedding network that projects text representations into a common space by calculating offset vectors between linguistic and non-linguistic information. Hazarika et al. [15] proposed modality-invariant and modality-specific representations to learn complementary information between modalities, reducing redundancy and merging a set of diverse information. Yu et al. [16] designed a label generation module based on a self-supervised learning strategy to capture consistency and differences between three modalities by jointly

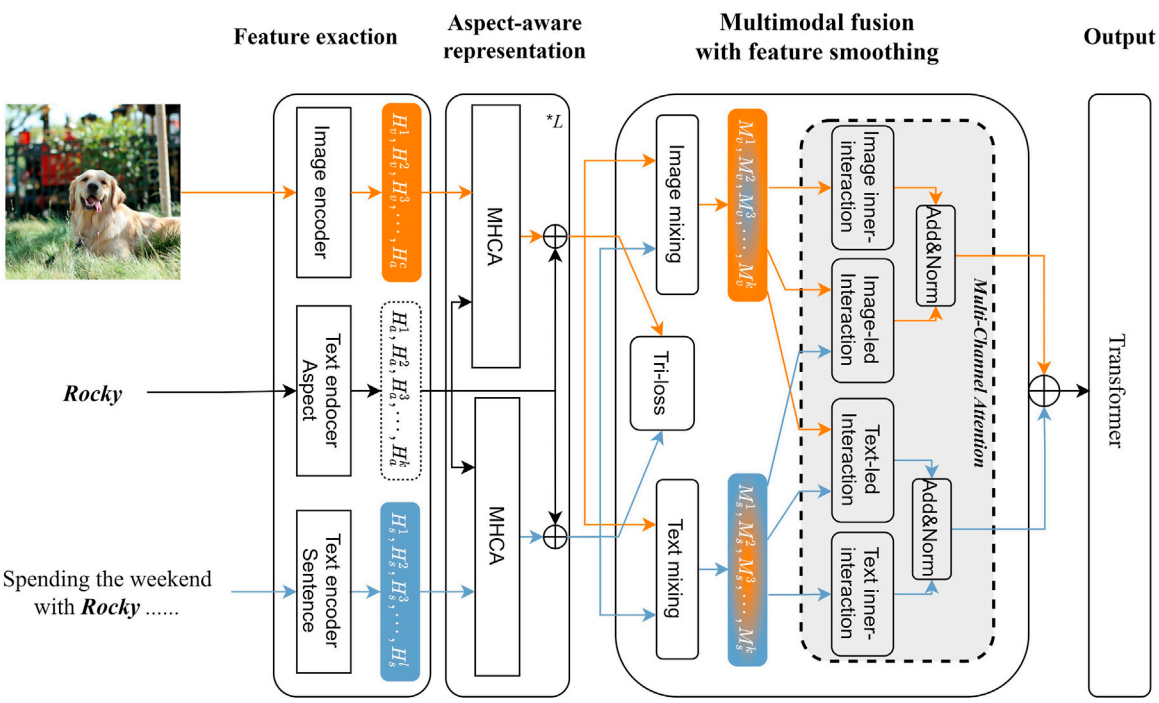


FIGURE 2
Overall framework.

learning unimodal and multimodal tasks. Effective unimodal representations can mitigate the impact of the semantic gap.

Multimodal fusion: For multimodal fusion, Zadeh et al. [17] proposed a tensor fusion network that obtains multimodal fusion representation by calculating outer products between all unimodal representations. Liu et al. [18] proposed an improved low-rank multimodal fusion network based on tensor fusion network, which uses low-rank tensors to reduce the computational complexity of tensor-based methods and achieve better performance. Zadeh et al. [19] proposed a memory fusion network that first models unimodal representations using LSTM, and then models intermodal interactions using Delta-memory Attention Network and Multi-view Gated Memory. Transformer structures are widely used to model interactions between modalities due to the success of Transformer-based models. Tsai et al. [20] used Directional Cross-Modal Attention modules to extend the standard Transformer network [21] for modeling unaligned multimodal language sequences. Wang et al. [22] used forward and backward translation from one modality to another and back to better fuse multimodal features. Modeling efficient interactions between different modalities can fully utilize information between modalities for multimodal emotional expression.

2.3 Multimodal aspect-based sentiment Classification (MABSC)

MABSC is the research combining ABSC and MSA. Similar to text aspect-based sentiment classification, different parts of the sentence and image play different roles in specific aspects, and attention mechanisms are widely used to obtain aspect-specific representations. Xu et al. [3] first used interactive attention to obtain aspect-specific unimodal representations, and then stacked several interactive attention mechanisms and memory hops to learn deep abstractions for multimodal data. Zhang et al. [4] used an aspect-sensitive memory network to capture intra-modal features, then designed a fusion discriminative matrix to learn interactions between different modalities. Inspired by the success of BERT-based models, Yu et al. [23] proposed a target-oriented multimodal BERT (TomBERT), which constructs a BERT-based structure to match the target text and target image and capture dynamics within and between modalities. Khan et al. [1] used a pre-trained transformer-based image captioning model to convert images into textual image captions, then fused information from both modalities by constructing sentence pairs and inputting the image caption, aspect, and original sentence into a BERT language model. Yu et al. [2] modeled pairwise interactions between inputs using an interactive transformer, and bridged the semantic gap between the two modalities by calculating the loss between the representations of the two modalities and the original context. Additionally, Huang et al. [24] constructed sequential cross-modal semantic graphs to fully extract the information contained in the image, and used an encoder-decoder model with a target prompt template to achieve MABSC task.

The importance of integrating image information into text information has been repeatedly proved in the research of MABSC. However, this integration invariably encounters the issue of semantic gaps between two modalities. Therefore, we focus on easing the semantic gap before integration.

3 Methodology

In this section, we first give the definition of multimodal aspect-based sentiment analysis task, and introduce the overall framework of the proposed model. Then, we present the details of each module of the proposed model.

3.1 Task definition

Given a set of multimodal dataset D, each sample $d \in D$ includes a context sentence t, an associated image i, a given aspect a, and a golden label y. Specifically, the sentence $t = (w_1, w_2, w_3, \ldots, w_m)$, where m is the length of the sentence. The given aspect is a subsequence of sentence t and is represented as $a = (w_x, w_{x+1}, \ldots, w_{x+n})$, where n is the length of the given aspect. As shown in Figure 2, this task is to take t, i and a as inputs to determine the sentiment polarity $y \in \{Positive, Neutral, Negative\}$ associated with the given aspect a.

3.2 Overview of the proposed model

The overall architecture of the model is shown in Figure 2, which consists of a feature extraction layer, an aspect-aware representation layer, a multimodal fusion layer with feature smoothing, and an output layer. We extract separate representations of the image, text and aspect in the feature extraction layer. In the aspect-aware representation layer, we mine the aspect-related representations of each modality with the guidance of the aspect. In the multimodal fusion layer, we use feature smoothing strategy and multi-channel attention to model the deep interaction between the two modalities. Finally, we obtain the sentiment polarities in the output layer.

3.3 Feature extraction layer

We utilize two different unimodal feature encoders to extract original representations of the text and image inputs.

3.3.1 Text encoder

The pre-training language model BERT [11] can capture advanced text representations. To distinguish sentence and aspect representations, we fine-tune two different pre-trained BERTs to encode sentence and aspect respectively. Specifically, for the input sentence, we add a special token [CLS] in front of the original sentence and a special token [SEP] in the back to form new tokens $\mathbf{I}_s \in \mathbb{R}^l$, and then input $\mathbf{I}_s$ to a pre-trained *BERT* to obtain the encoded sentence representation $\mathbf{h}_s$, as follows:

$$\mathbf{h}_s = BERT(\mathbf{I}_s) \tag{1}$$

where $\mathbf{h}_s \in \mathbb{R}^{l*d_t}$ is the obtained sentence representation, d_t is the hidden dimension.

Similarly, for a given aspect, we add the special tokens [CLS] and [SEP] to form tokens $\mathbf{I}_a \in \mathbb{R}^k$, and then input $\mathbf{I}_a$ into another pre-trained *BERT* to obtain the encoded aspect representation $\mathbf{h}_a$, as follows:

$$\mathbf{h}_a = BERT(\mathbf{I}_a) \tag{2}$$

where $\mathbf{h}_a \in \mathbb{R}^{k*d_t}$ is the obtained aspect representation.

After obtaining the sentence and aspect representation, we use the linear layer to map their hidden dimension to the same dimension d_h for the subsequent interaction:

$$\mathbf{H}_s = \mathbf{W}_1\mathbf{h}_s + \mathbf{b}_1 \tag{3}$$

$$\mathbf{H}_a = \mathbf{W}_2\mathbf{h}_a + \mathbf{b}_2 \tag{4}$$

where $\mathbf{H}_s \in \mathbb{R}^{l*d_h}$ and $\mathbf{H}_a \in \mathbb{R}^{k*d_h}$.

3.3.2 Image encoder

Different from coarse grained sentiment analysis tasks, MABSC should focus on aspect-related information to determine the sentiment polarity. We use the object detection model Faster R-CNN [25] to extract aspect-level features of images. Specifically, we input the image i into a pre-trained *Faster R-CNN* model to obtain the candidate regions in the image, and retain the features with the highest confidence as image features:

$$\mathbf{h}_i = FasterR-CNN(i) \tag{5}$$

where $\mathbf{h}_i \in \mathbb{R}^{c*d_v}$ is the obtained image representation, c denotes the number of image regions retained, and d_v is the hidden dimension of *Faster R-CNN*.

Then we use a linear layer to map the hidden dimension of image representation to d_h:

$$\mathbf{H}_i = \mathbf{W}_3\mathbf{h}_i + \mathbf{b}_3 \tag{6}$$

where $\mathbf{H}_i \in \mathbb{R}^{c*d_h}$.

We obtain the final image representation by a multi-head self attention (*MHSA*) [21] to pay more attention to the important image regions:

$$\mathbf{H}_v = MHSA(\mathbf{H}_i) \tag{7}$$

where $\mathbf{H}_v \in \mathbb{R}^{c*d_h}$.

3.4 Aspect-aware representation layer

After obtaining the initial sentence representation and image representation, we need to further interact them with the aspect representation to focus on aspect-related information. We adopt an interactive attention mechanism to enable interaction between the aspect representation and unimodal representation, and retain more aspect representations through residual connections. Specifically, we use the aspect representation as the query, and the sentence representation as the key-value in the multi-head cross attention (*MHCA*) [21], to generate the aspect-sentence representation, as follows:

$$\mathbf{R}_s = MHCA(\mathbf{H}_a, \mathbf{H}_s) \tag{8}$$

where $\mathbf{R}_s \in \mathbb{R}^{k*d_h}$.

Then we add the aspect-sentence representation and the aspect representation, and perform one layer normalization (*LN*) to obtain the one-layer aspect-aware text representation:

$$\mathbf{A}_s = LN(\mathbf{R}_s + \mathbf{H}_a) \tag{9}$$

where $\mathbf{A}_s \in \mathbb{R}^{k*d_h}$.

Finally, we stack l layers of the aspect-aware layer to learn the deep interaction of aspect and text, as follows:

TABLE 1 Dataset statistics.

	Twitter2015			Twitter2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	417	573
Negative	368	149	113	416	144	168
Total Samples	3179	1122	1037	3562	1176	1234
Avg Aspect	1.348	1.336	1.354	1.410	1.439	1.450
Avg Length	16.72	16.74	17.05	16.21	16.37	16.38
Max Length	35	40	36	39	31	38
Total Sentence	2101	727	674	1746	577	587

TABLE 2 The hyperparameter Setting.

	Twitter2015	Twitter2017
Learning rate	2e-5	4e-5
Warm up step	37	35
l	2	1
W_{mix}	0.85	0.85
(α_1, α_2, α_3)	(1,1,0.5)	(1,1,0.5)
λ	4e-3	4e-3
Batch size	32	32
Attension heads	8	8
Attention dimension	512	512

$$\mathbf{A}_{s(l)} = LN\left(MHCA\left(\mathbf{H}_a, \mathbf{A}_{s(l-1)}\right) + \mathbf{H}_a\right) \tag{10}$$

where $\mathbf{A}_{s(l)} \in \mathbb{R}^{k*d_h}$ is the final aspect-aware text representation, and l is the number of stacked layers.

For the image representation, we input it together with the aspect representation into the similar aspect-aware layer to obtain the aspect-aware image representation, as follows:

$$\mathbf{A}_{v(l)} = LN\left(MHCA\left(\mathbf{H}_a, \mathbf{A}_{v(l-1)}\right) + \mathbf{H}_a\right) \tag{11}$$

where $\mathbf{A}_{v(l)} \in \mathbb{R}^{k*d_h}$.

3.5 Multimodal fusion layer with feature smoothing

After obtaining the aspect-aware representations of two modalities, we propose a multimodal fusion layer with feature smoothing to combine information from different modalities. Firstly, to relieve the semantic gap between the two modalities, a feature smoothing strategy is used to smooth the aspect-aware representations of the two modalities. Then we use a multi-channel attention interaction network to achieve deep interaction between the two modal representations.

3.5.1 Feature smoothing

We integrate the partial representation of one modality into the representation of another modality via a feature-level mixing approach, and obtain two smoothed unimodal representations, as follows:

$$\mathbf{M}_s = W_{mix}*\mathbf{A}_{s(l)} + (1 - W_{mix})*\mathbf{A}_{v(l)} \tag{12}$$

$$\mathbf{M}_v = W_{mix}*\mathbf{A}_{v(l)} + (1 - W_{mix})*\mathbf{A}_{s(l)} \tag{13}$$

where W_{mix} is a hyperparameter. The obtained $\mathbf{M}_s$ and $\mathbf{M}_v$ are the smoothed text and image representations, respectively. We will use these smoothed representations for further interaction.

In addition, we use the average representation of the two modalities as an anchor, to bridge the semantic gap between the two modalities via the constraint of the mean square error Tri-loss:

$$\mathbf{A}_l = MEAN\left(\mathbf{A}_{s(l)}, \mathbf{A}_{v(l)}\right) \tag{14}$$

$$L^{tri} = \alpha_1*MSE\left(\mathbf{A}_{s(l)}, \mathbf{A}_l\right) + \alpha_2*MSE\left(\mathbf{A}_{v(l)}, \mathbf{A}_l\right) + \alpha_3*MSE\left(\mathbf{A}_{s(l)}, \mathbf{A}_{v(l)}\right) \tag{15}$$

where the *MEAN* operator refers to averaging values of each dimension in the two tensors. *MSE* is mean square error loss, and (α_1, α_2, α_3) are hyperparameters. The above loss will be added to the main loss to guide the training of the model parameters.

3.5.2 Multi-channel attention-based interaction

In order to effectively utilize the complementary information between modalities to enhance the expression of sentiment, we propose a multi-channel attention interaction network (MCA) including four channels, named text self-attention, text-led multimodal attention, image self-attention and image-led multimodal attention channels respectively.

In the text self-attention channel, we use a multi-head self attention to process the smoothed text representation acquired in the preceding stage and obtain the text inner-interaction representation, denoted as $\mathbf{CS}_s \in \mathbb{R}^{k*d_h}$:

$$\mathbf{CS}_s = MHSA\left(\mathbf{M}_s\right) \tag{16}$$

In the text-led multimodal attention channel, we take the smoothed text representation as the query and the smoothed image representation as the key-value, and sent them to a multi-head interactive attention network, to obtain the text-led inter-interaction representation, denoted as $\mathbf{CC}_s \in \mathbb{R}^{k*d_h}$:

$$\mathbf{CC}_s = MHCA\left(\mathbf{M}_s, \mathbf{M}_v\right) \tag{17}$$

Final, we add up the representations of the two channels and normalize it to obtain the text-led multimodal representation $\mathbf{F}_s \in \mathbb{R}^{k*d_h}$:

$$\mathbf{F}_s = LN\left(\mathbf{CS}_s + \mathbf{CC}_s\right) \tag{18}$$

Similarly, following the same procedure as the two text channels above, we feed the smoothed image representation into the two image channels to obtain the image inner-interaction representation and the image-led inter-interaction representation. We then add and normalize them to obtain the image-led multimodal representation $\mathbf{F}_v \in \mathbb{R}^{k*d_h}$.

TABLE 3 Comparison of our method and baseline Macro-F1.

Modality	Method	Twitter2015	Twitter2017
Visual	Res-Aspect	46.58	54.01
	FasterRCNN-Aspect	37.71	54.71
Text	IAN [9]	63.32	63.32
	MGAN [10]	64.21	61.46
	BERT [11]	70.01	66.15
Text + Visual	Res-BERT	71.46	66.89
	Faster R-CNN-BERT	70.85	66.21
	TomBERT (ResNet) [23]	71.75	68.04
	TomBERT (FasterR-CNN) [2]	72.95	68.49
	ModalNet [4]	72.50	69.19
	IFNRA [27]	71.79	69.48
	MSFNet (Ours)	74.46	69.67

TABLE 4 Ablation study of feature-level mixing (Macro-F1).

Method	Twitter2015	Twitter2017
MSFNet (Ours)	74.46	69.67
w/o feature mixing	72.83	68.23
w/o Text mixing	73.88	68.91
w/o Image mixing	73.86	68.92

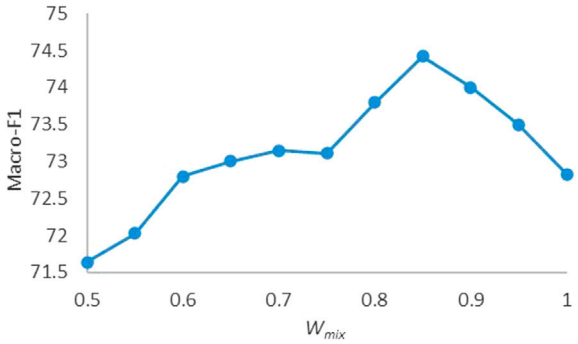


FIGURE 3
Performance of different weight parameter W_{mix}.

After obtaining the two representations $\mathbf{F}_s$ and $\mathbf{F}_v$, we concatenate them and send it to a transformer and a average pooling, to get the final multimodal sentiment representation $\mathbf{H}_m \in \mathbb{R}^{d_h}$:

$$\mathbf{F}_m = [\mathbf{F}_s; \mathbf{F}_v] \tag{19}$$

$$\mathbf{H}_m = averagepooling\left(Transformer\left(\mathbf{F}_m\right)\right) \tag{20}$$

3.6 Output layer

We send the multimodal sentiment representation $\mathbf{H}_m$ to a fully connected layer and a softmax layer to obtain the classification result:

$$p\left(y|\mathbf{H}_m\right) = softmax\left(\mathbf{W}_c\mathbf{H}_m + \mathbf{b}_c\right) \tag{21}$$

where $\mathbf{W}_c \in \mathbb{R}^{r*d_h}$ and $\mathbf{b}_c \in \mathbb{R}^r$ are learnable parameters, $y \in \mathbb{R}^r$ is the probability distribution of sentiment polarity, r is number of classes.

The loss function of the model is as follows:

$$L = -\frac{1}{N}\sum_{i}^{N}\left(\sum_{j}^{r} g_{ij}\log p\left(y_{ij}|\mathbf{H}_m\right) - \lambda L_i^{tri}\right) \tag{22}$$

where g_{ij} is the golden label, λ is a hyperparameter.

4 Experimental

In this section, we conducted comprehensive experiments on the proposed overall model and its individual modules.

4.1 Experiment setting

Datasets: We adopt two standard datasets Twitter15 and Twitter17 to evaluate the performance of our model. Twitter15 and Twitter17 datasets contain multimodal tweets published on Twitter between 2014–2015 and 2016–2017 respectively. These datasets were originally annotated the given aspect by Zhang et al. [26] for the Multimodal Named Entity Recognition (MNER) task, and then Yu et al. [23] annotated the sentiment polarity of each given aspect for the MABSA task. The datasets provide tweet text, tweet image, aspect and the sentiment polarity of the given aspect. The specific data statistics are shown in Table 1.

Evaluation Metrics: To measure the performance of different approaches, we use Macro-F1 as evaluation metrics, as follows:

$$Macro - F1 = \frac{1}{r}\sum_{i=1}^{r} F1_i \tag{23}$$

$$F1_i = \frac{2*P_i*R_i}{P_i + R_i} \tag{24}$$

where $F1_i$ is the *f1-score* of class i, P_i and R_i are the precision and recall of class i, and r is the number of classes.

Implement Details: For text input, we leverage the pre-trained BERT [11] model to encode the text. For image input, we utilized the Faster R-CNN structure proposed by Anderson et al. [25] and used a pre-trained Faster R-CNN model to extract region features of the image. We fix all the hyper-parameters after tuning them on the development set. The specific hyperparameter settings are shown in Table 2. We implemented all models in the PyTorch framework and ran experiments on RTX3090 GPU.

4.2 Baseline

In this section, we use the following methods as baselines to compare with our model.

TABLE 5 Ablation study of Tri-loss (Macro-F1).

Method	Twitter2015	Twitter2017
MSFNet (Ours)	74.46	69.67
Rep anchor of Tri-loss	73.71	69.02
w/o Tri-loss	73.08	68.36

TABLE 6 Ablation study of MCA (Macro-F1).

Method	Twitter2015	Twitter2017
MSFNet (Ours)	74.46	69.67
w/o MCA	71.21	67.47
Rep MCA to CMT	70.73	67.24

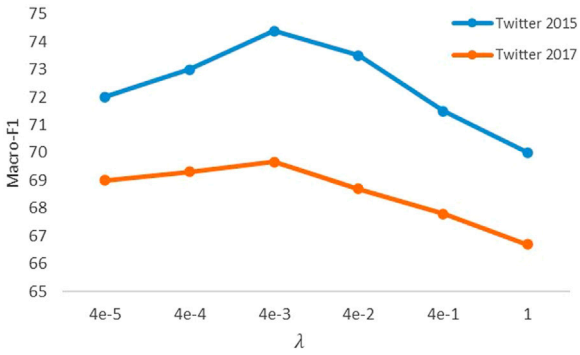


FIGURE 4
Performance of different weight parameter λ.

- Res-Aspect: ResNet and BERT are used to extract image and aspect features respectively, and an attention layer is used to obtain multimodal representation.
- Faster R-CNN-Aspect: Another baseline is similar to Res-Aspect, but image features are extracted by Faster R-CNN.
- IAN [9]: Capturing the interaction between aspect and context with bidirectional interactive attention.
- MGAN [10]: Based on IAN, a fine-grained attention is further proposed for interaction.
- BERT [11]: Sentence pairs constructed by context and aspect are fed into pre-trained BERT for sentiment classification.
- Res-BERT: The context and aspect are input as sentence pairs into a pre-trained BERT model to obtain text features. The image features are extracted by ResNet. And then modeling multimodal interaction using attention.
- FasterR-CNN-BERT: Another baseline is similar to Res-BERT, but image features are extracted by Faster R-CNN.
- TomBERT (ResNet) [23]: A target-oriented multimodal BERT architecture that utilizes ResNet for image representation, and leverages multiple BERT structures for text feature extraction, image aspect interaction, and multimodal interaction.
- TomBERT (Faster R-CNN) [2]: Same structure as TomBERT (ResNet), but the image representation is obtained by Faster R-CNN.
- ModalNet [4]: Use aspect-sensitive memory network to perform aspect-sensitive fusion of two modalities, and construct a fusion discriminant matrix to obtain multimodal sentiment representation.
- IFNRA [27]: Use GRU to achieve image denoising and multimodal fusion. And a decoder with recurrent attention is designed to gradually learn aspect-specific sentiment features.

4.3 Main result

Table 3 shows the performance of different methods on the twitter2015 and twitter2017 datasets. The following observations can be drawn: (1) Our model has achieved the best performance on the two datasets, which are respectively improved by 1.96% and 0.19% compared with the second best model. This illustrates that our proposed multimodal fusion method is effective and has obvious advantages. (2) Sufficient multimodal fusion can effectively improve classification performance. For example, both TomBERT (Faster R-CNN) and Faster R-CNN-BERT use Faster R-CNN to extract regional features, but the latter performs much worse than the former because it only performs simple multimodal fusion. Similarly, for the models that use ResNet to extract image features, TomBERT (ResNet) shows better performance than Res-BERT, but it is still not as good as ModalNet. Our proposed method has significant advantages when compared to ModalNet. The latter focuses on multimodal fusion without considering the semantic gap of multimodal features. Our proposed model performs feature smoothing before multimodal fusion, which enables deeper interactions and achieves better performance. (3) Using the regional features extracted by FasterR-CNN can help the model focus on the object-level information in images. However, if the model cannot obtain information enabling to expressing sentiments from the image representation via a good image-text interaction method, using FasterR-CNN may result in performance degradation. This conclusion can be drawn from comparing Res-BERT and Faster R-CNN-BERT, as well as Res-Aspect and Faster R-CNN-Aspect. (4) The performance of image-based methods is much lower than that of text-based methods among the unimodal-based methods. This is mainly because the given aspect is a subsequence in the initial sentence. If image information is considered alone, it may introduce some noises that have nothing to do with the given aspect, resulting in wrong classification.

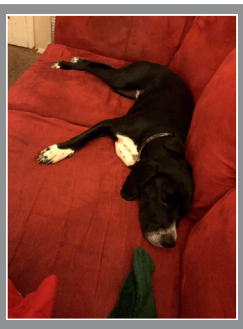
4.4 Ablation study of multimodal fusion layer

In this section, we conduct ablation studies to verify the effectiveness of multimodal fusion layer with feature smoothing.

4.4.1 Feature-level mixing

To test the effect of feature-level mixing, we feed the unprocessed aspect-aware representations into the multi-channel attention interaction network instead of smoothed representations. The results are shown in Table 4.

TABLE 7 Comparison between predicted results and golden labels for several representative samples on Bert, Faster R-CNN-BERT and MSFNet (Ours), respectively.

Image		
Text	(a) ***Charlie*** is decidedly not excited about @ ***ussoccer_ynt*** at 4 am. #U20WC	(b) The final chapter of the fairytale—***Leicester*** gear up for historic ***Premier League*** title
Golden Label	(Charlie, Negative)	(Leicester, Positive)
	(ussoccer_ynt, Neutral)	(Premier League, Neutral)
Bert	(Charlie, Neutral) ✗	(Leicester, Neutral) ✗
	(ussoccer_ynt, Neutral) ✓	(Premier League, Neutral) ✓
FasterR-CNN-BERT	(Charlie, Negative) ✓	(Leicester, Neutral) ✗
	(ussoccer_ynt, Neutral) ✓	(Premier League, Neutral) ✓
MSFNet (Ours)	(Charlie, Negative) ✓	(Leicester, Positive) ✓
	(ussoccer_ynt, Neutral) ✓	(Premier League, Neutral) ✓

It can be seen that if the unprocessed aspect-aware features of one modality are used to interact with the smooth features of another modality, the Macro-F1 of the twitter15 and twitter17 datasets drop by about 0.5% and 0.7%, respectively, compared to the full model. If all four interaction channels use unmixed features, the Macro-F1 drops by more than 1.44% on the two datasets. The above results further shows that feature smoothing before image-text interaction can better achieve multimodal fusion and improve classification performance.

In feature-level mixing, we set a hyperparameter to control the smoothing weight. Figure 3 shows the impact of different weight on the model performance of the twitter15 dataset. Setting the hyperparameter to 1 means that the two modalities do not perform feature smoothing, while setting to 0.8–0.95 means that we take one modality as the dominant information and incorporate a little information from another modality. It can be seen that when the hyperparameter is set to 0.8–0.95, the model can obtain better results than 1 or less than 0.75. This may be because, when feature smoothing is not performed, the semantic gap between modalities will make subsequent interactions insufficient. In addition, if we incorporate too much information from another modality, the dominant modal will lose its own representational ability. The best performance is achieved when the dominant modal feature introduces around 15% of the other modal feature.

4.4.2 Tri-loss

In Table 5 we report the ablation study of the Tri-loss. It can be seen that the performance drops sharply after the removal of Tri-loss, which illustrates the effectiveness of reducing the semantic distance between the two modalities via the constraint of Tri-loss. What's more, if we use the initial aspect representation instead of the average representation of the two modals as the anchor in the Tri-loss, the performance decreases too. The reason may be that the model would learn from the lower-level aspect representation if using the initial aspect representation as the anchor after aspect-aware fusion, which is ineffective.

We adjusted the weight parameter λ of Tri-loss in the total loss to observe its effection. It can be seen from Figure 4 that the model achieves the best performance when λ is 4e-3, while assigning too large or small weight leads to a decrease in the final performance. This illustrates that using appropriate constraints of Tri-loss can benefit the model.

4.4.3 Multi-channel attention

We verified the effectiveness of the multi-channel attention-based interaction (MCA) by deleting it or replacing it with the Cross-Modal Transformer (CMT) [20]. As can be seen in Table 6, the performance decreases by 3.25% and 2.23% on the two datasets respectively after removing the module, which illustrates the necessity of performing deep image-text fusion. Furthermore, the performance decreases by 3.73% and 2.43% on the two datasets after replacing MCA with CMT, which fully illustrates the effectiveness of our proposed MCA module.

4.5 Case study

In this section, we choose two representative samples to compare the prediction results of our model with the two

baselines. Firstly, in Table 7, BERT predicted the sentiment polarity of the aspect {Charlie} incorrectly, which could be due to BERT only predicts based on text content and cannot recognize the negative sentiment expressed by the corresponding aspect in the image. In addition, the model Faster R-CNN-BERT, which also uses Faster R-CNN to capture image object-level features, made wrong predictions for the aspect Leicester in Table 7, while our model made correct predictions. It may be due to our excellent fusion network that enables our model to accurately capture the positive emotions expressed by waving flag in the image.

5 Conclusion

In this paper, we propose a MABSC model based on a multimodal feature smoothing fusion network. We extracts aspect-aware representations of text and image modals at first. Then, we introduce a feature smoothing strategy to get smoothed representations, which are sent to the proposed multi-channel attention-based network for image-text information interaction. By this process, the comprehensive aspect-level sentiment representation is obtained for better classification. Experiments demonstrate that the model achieves better performance than the other baselines on the two datasets. The ablation experiments further demonstrate the effectiveness of the various modules of the model. In the future work, we will further consider how to align aspect-related information in image and text content, given that MABSC task requires to focus on fine-grained information in image and text.

Data availability statement

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

Author contributions

YX: Conceptualization, Methodology, Validation, Formal analysis, Writing—Review and Editing. YC: Software, Investigation, Writing—Original Draft. JG: Conceptualization, Methodology, Software, Writing—Review and Editing. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work is supported by the National Natural Science Foundation of China (Grant Nos 62162037, 62266027, U21B2027, 62266028), General projects of basic research in Yunnan Province (Grant Nos 202001AT070047, 202001AT070046, 202301AT070444), Kunming University of Science and Technology "double first-class" joint project (Grant No. 202201BE070001-021).

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. Khan Z, Fu Y. Exploiting bert for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM International Conference on Multimedia; New York, NY, USA. New York, NY: Association for Computing Machinery (2021). p. 3034–42. MM '21. doi:10.1145/3474085.3475692

2. Yu J, Chen K, Xia R, Wang Y, Feng K, Wan T, et al. Comprehensive comparisons of ocular biometry: A network-based big data analysis. *IEEE Trans Affective Comput* (2022) 10:1. doi:10.1186/s40662-022-00320-3

3. Xu N, Mao W, Chen G. Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence; July 2019. Palo Alto, CA: AAAI Press (2019). AAAI'19/IAAI'19/EAAI'19. doi:10.1609/aaai.v33i01.3301371

4. Zhang Z, Wang Z, Li X, Liu N, Guo B, Yu Z. Modalnet: An aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web* (2021) 24:1957–74. doi:10.1007/s11280-021-00955-7

5. Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2nd International Conference on Knowledge Capture; October 2003; New York, NY, USA. New York, NY: Association for Computing Machinery (2003). p. 70–7. K-CAP '03. doi:10.1145/945645.945658

6. 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 (SemEval 2014); January 2014; Dublin, Ireland. Dublin, Ireland: Association for Computational Linguistics (2014). p. 437–42. doi:10.3115/v1/S14-2076

7. Tang D, Qin B, Feng X, Liu T. Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers; Osaka, Japan. The COLING 2016 Organizing Committee (2016). p. 3298–307.

8. Nguyen HT, Le Nguyen M. Effective attention networks for aspect-level sentiment classification. In: Proceedings of the 2018 10th International Conference on Knowledge and Systems Engineering (KSE); November 2018; Ho Chi Minh City, Vietnam. IEEE (2018). p. 25–30. doi:10.1109/KSE.2018.8573324

9. Ma D, Li S, Zhang X, Wang H. Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence IJCAI- 17; August 2017. Palo Alto, CA: Association for Computing Machinery (2017). p. 4068–74. doi:10.24963/ijcai.2017/568

10. Fan F, Feng Y, Zhao D. Multi-grained attention network for aspect-level sentiment classification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Brussels, Belgium. Brussels, Belgium: Association for Computational Linguistics (2018). p. 3433–42. doi:10.18653/v1/D18-1380

11. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Minneapolis, Minnesota. Minneapolis, Minnesota: Association for Computational Linguistics (2019). p. 4171–86. doi:10.18653/v1/N19-1423

12. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2019; Minneapolis, Minnesota. Minneapolis, MinnesotaAssociation for Computational Linguistics (2019). p. 380–5. doi:10.18653/v1/N19-1035

13. Morency LP, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th International Conference on Multimodal Interfaces; November 2011; New York, NY, USA. New York, NY: Association for Computing Machinery (2011). p. 169–76. ICMI '11. doi:10.1145/2070481.2070509

14. Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency LP. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence; July 2019. Palo Alto, CA: AAAI Press (2019). AAAI'19/IAAI'19/EAAI'19. doi:10.1609/aaai.v33i01.33017216

15. Hazarika D, Zimmermann R, Poria S. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia; New York, NY, USA. New York, NY: Association for Computing Machinery (2020). p. 1122–31. MM '20. doi:10.1145/3394171.3413678

16. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proc AAAI Conf Artif Intelligence* (2021) 35:10790–7. doi:10.1609/aaai.v35i12.17289

17. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; September 2017; Copenhagen, Denmark. Copenhagen, Denmark: Association for Computational Linguistics (2017). p. 1103–14. doi:10.18653/v1/D17-1115

18. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Bagher Zadeh A, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 2018. Melbourne, Australia: Association for Computational Linguistics (2018). p. 2247–56. doi:10.18653/v1/P18-1209

19. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP. Memory fusion network for multi-view sequential learning. *Proc AAAI Conf Artif Intelligence* (2018) 32. doi:10.1609/aaai.v32i1.12021

20. Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy. Palo Alto, CA: Association for Computational Linguistics (2019). p. 6558–69. doi:10.18653/v1/P19-1656

21. 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.

22. Wang Z, Wan Z, Wan X. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In: Proceedings of The Web Conference 2020; September 2020; New York, NY, USA. New York, NY: Association for Computing Machinery (2020). WWW '20, 2514–2520. doi:10.1145/3366423.3380000

23. Yu J, Jiang J. Adapting bert for target-oriented multimodal sentiment classification. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. San Mateo, CA: International Joint Conferences on Artificial Intelligence Organization (2019). p. 5408–14. doi:10.24963/ijcai.2019/751

24. Huang Y, Chen Z, Zhang W, Chen J, Pan JZ, Yao Z, et al. *Aspect-based sentiment classification with sequential cross-modal semantic graph* (2022). ArXiv abs/2208.09417.

25. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2018). p. 6077–86. doi:10.1109/CVPR.2018.00636

26. Zhang Q, Fu J, Liu X, Huang X. Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA: AAAI Press (2018). AAAI'18/IAAI'18/EAAI'18.

27. Wang J, Wang Q, Wen Z, Liang X, Xu R. Interactive fusion network with recurrent attention for multimodal aspect-based sentiment analysis Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part III (Berlin, Heidelberg: Springer-Verlag) (2022), 298–309. doi:10.1007/978-3-031-20503-3_24

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 19 May 2023
DOI 10.3389/fphy.2023.1192412

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Hao Tang,
Nanjing University of Science and Technology, China
Kangjian He,
Yunnan University, China
Puhong Duan,
Hunan University, China

*CORRESPONDENCE
Fan Li,
292789049@qq.com

RECEIVED 23 March 2023
ACCEPTED 18 April 2023
PUBLISHED 19 May 2023

CITATION
Zhou H, Li F, Tian X and Huang Y (2023), Feature semantic alignment and information supplement for Text-based person search.
Front. Phys. 11:1192412.
doi: 10.3389/fphy.2023.1192412

Feature semantic alignment and information supplement for Text-based person search

Hang Zhou[1], Fan Li *, Xuening Tian [2] and Yuling Huang[3]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [2]LTH Engineering College at Campus Helsingborg, Lund University, Lund, Sweden, [3]School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

The goal of person text-image matching is to retrieve images of specific pedestrians using natural language. Although a lot of research results have been achieved in persona text-image matching, existing methods still face two challenges. First,due to the ambiguous semantic information in the features, aligning the textual features with their corresponding image features is always tricky. Second, the absence of semantic information in each local feature of pedestrians poses a significant challenge to the network in extracting robust features that match both modalities. To address these issues, we propose a model for explicit semantic feature extraction and effective information supplement. On the one hand, by attaching the textual and image features with consistent and clear semantic information, the course-grained alignment between the textual and corresponding image features is achieved. On the other hand, an information supplement network is proposed, which captures the relationships between local features of each modality and supplements them to obtain more complete local features with semantic information. In the end, the local features are then concatenated to a comprehensive global feature, which capable of precise alignment of the textual and described image features. We did extensive experiments on CUHK-PEDES dataset and RSTPReid dataset, the experimental results show that our method has better performance. Additionally, the ablation experiment also proved the effectiveness of each module designed in this paper.

KEYWORDS

cross-modal retrieval, neural network, Text-based person search, deep learning, Text-based image retrieval

1 Introduction

Person text-image matching method has been proposed in order to deal with special cases. For example, if a child is missing in an amusement park, parents can quickly find the area where the child is located in the surveillance equipment by describing the child's appearance. This technique uses the textual description of the pedestrian's appearance provided to retrieve the target pedestrian image. Compared with person re-identificationLi et al. [1]; Zhang et al. [2]; Wang et al. [3]; Zhang et al. [4]; Li et al. [5], it is not limited to the need of pedestrian images as a query condition, so it compensates for the disadvantages of using pedestrian re-identification techniques in the presence of surveillance blind spotsZhu et al. [6]; Li et al. [7]; Lingli et al. [8]; Li et al. [9]. Therefore the technique has practical value.

In this task, ensuring the consistency of text semantic information and image semantic information is one of the factors that affect the retrieval performance. In recent years, many

effective methods of feature semantic consistency have been proposed. These methods can be roughly divided into methods based on local relationship correspondence Ding et al. [10]; Zhang et al. [11]; Liu et al. [12]; Zheng et al. [13]; Chen et al. [14], etc.), methods based on external knowledge Jing et al. [15]; Wang et al. [16]; Aggarwal et al. [17]; Wang et al. [18], methods based on similarity measurement Niu et al. [19]; Gao et al. [20], and methods based on multi-head attention mechanism Wang et al. [21]; Li et al. [22]. The method based on local relation correspondence often achieves local alignment of text features and image features through specific functional relations. In the method based on external knowledge, human body semantics Wang et al. [16], pedestrian posture Jing et al. [15] and pedestrian attributes Li et al. [23]; Wang et al. [18] are often used as auxiliary information for text and visual feature alignment. The method based on similarity measure measures the similarity between noun phrases and local patches in the image, and determines the relationship between them according to the predicted weight. Compared with the attention mechanism used in the method based on local relationship correspondence, the method based on multi-head attention mechanism usually assigns different semantics to each head to align the heads with the same semantics.

Although many effective methods have been proposed, there are still some problems that have not been effectively solved. On the one hand, features alignment based on local relationships always faces the problem of vague semantic information in local features. This is because different images of the same person may include spatial discrepancy and different sentences describing the same person may have differences in the order of expression and logic. On the other hand, some special semantic information in the text, such as "Coat," corresponds to multiple body parts, such as the upper and lower body, in the image. Therefore, those various semantics are not entirely independent, and there should be correlations between multiple different semantics. Previous methods for extracting local features have not put these correlations in their consideration, resulting in the loss of some semantic information in the local features and presents a challenge in building up local correspondences between text and images. Although attention-based methods Zheng et al. [13]; Liu et al. [12]; Gao et al. [20] can effectively alleviate this problem, they require high computational cost and loss of efficiency.

This paper proposes a method for explicit semantic feature extraction and an information supplement network to address the challenge of aligning textual and image features of pedestrians. The relationships between key information in the sentence have been fully considered. On the one hand, we starts with a class token obtained from the transformer, and predicts the features that correspond to the local regions of the pedestrian image from the global features with relationship embedding, thus obtaining local features that are roughly aligned and have clear semantics. On the other hand, the information supplement network is proposed to adaptively probe the relationships between local features, and such relationships are used to fuse out semantically well-informed local features. In the end, the local features with improved semantic information are precisely aligned and concatenated in a certain order on the channel to form globally aligned features with comprehensive semantics.

Our research contributions are as follows

- In this paper, we propose a explicit semantic feature extraction method. We use local features of pedestrian image with clear semantics to guide text feature extraction with fuzzy semantics, and achieve a rough alignment between local features of text and local features of images.
- To address the challenge of semantics loss in local features, this paper proposes an effective information supplement network to complement the missing information.

2 Related works

2.1 Text-based image retrieval

Text-based image retrieval Liu et al. [24] is a technique that uses natural language to retrieve specific images. It differs from the single-modal task Tang et al. [25,26]; Zha et al. [27]; Li et al. [28] in that it requires overcoming greater modal differences. Depending on the testing process, we can divide these methods into modal interaction methods Gao et al. [20] and modal non-interaction methods Chen et al. [14]; Liu et al. [24]. The modal interaction methods often requires each text feature with all image features to derive results through a complex cross-modal attention mechanism, which undoubtedly increases the time cost and it's hard to deploy in a real world scenario. The modal non-interaction methods can extract image features or text features separately and does not require two modality features for cross-learning, saving time overhead. Therefore, it can be used for large-scale text image retrieval tasks. However, these methods do not take into account the impact of semantic clarity and missing feature information on cross-modality matching.

2.2 Person Text-image matching

Person text-image matching faces different problems than Text-based Image Retrieval. In a Text-based Image Retrieval task, an image usually contains multiple objects, and the model design often needs to consider the association between objects. In the person text-image matching task, there is usually only one pedestrian object, so the model design needs to consider extracting fine-grained features. Therefore, by comparison, person text-image matching is more challenging. Li et al.Li et al. [29] first proposed the person text-image matching task and successfully completed the task using recurrent neural network with gated neural attention mechanism. Meanwhile, a large-scale person description dataset named the CUHK person description dataset was constructed. Because there are too many defects in the first proposed method. Subsequently, Li et al. Li et al. [30] proposed an identity aware two-stage network. The network extracts robustness features through two steps.

In recent years researchers have proposed a variety of methods, which can be broadly classified into the following three categories: similarity relation metrics based methods Niu et al. [19]; Gao et al. [20]; Li et al. [30], external knowledge assistance based methods Wang et al. [16]; Jing et al. [15]; Aggarwal et al. [17]; Wang et al. [18], and multi-granularity relational correspondence based feature alignment methods Ding et al. [10]; Zhang et al. [11]; Liu et al. [12]; Zheng et al. [13]; Chen et al. [14]; Wang et al. [21]. Similarity relation metrics based methods use the similarity between text features and image features as the relationship between them to obtain robust features. Then, during testing, it also requires each text

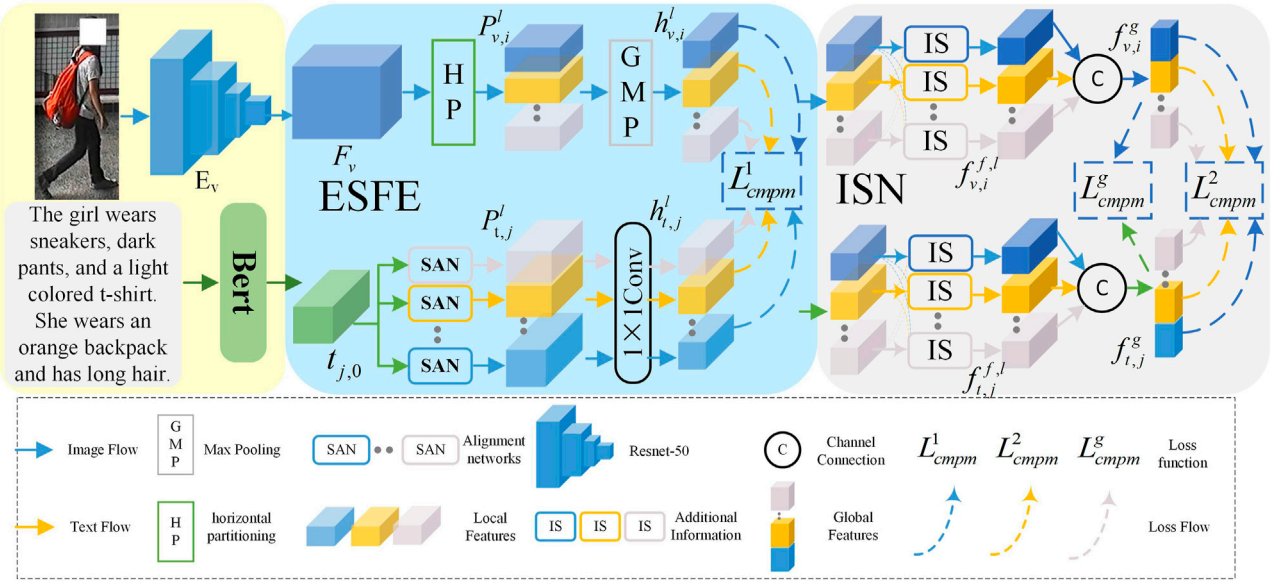


FIGURE 1
Overall framework of the proposed method.

to do the same operation with all images, which greatly reduces the testing efficiency. External knowledge assistance based methods need to construct external knowledge in advance to assist the model in extracting features. However, the model performance is highly dependent on the external knowledge, and the performance also depends on how well the external knowledge is constructed. Multi-granularity relational correspondence based feature alignment methods usually align the features of each granularity directly, which reduces the model performance without explicit semantic. This reduces the performance of the model without explicit semantics. In contrast to the above methods. To accomplish semantic alignment between textual and pedestrian image features, this study proposes obtaining distinct local features by projecting the global features of text onto the local feature space of the corresponding pedestrian image through a nonlinear mapping mechanism. Subsequently, the information supplementation network complements local feature information to achieve refined alignment of local features with comprehensive information. Utilizing these aligned local features, global features with coherent semantic information are then constructed.

3 Proposed method

3.1 Overview

The technical framework of this paper consists of two main parts: Explicit Semantic Feature Extraction (ESFE) and Information Supplementation Network (ISN), as shown in Figure 1. ESFE guides the image features with clear semantics to align with the vague semantic text features, thereby achieving semantic alignment and laying a solid foundation for downstream tasks. ISN is responsible for establishing the relationships between various local features and fusing them based on these relationships to eliminate the incompleteness of the features and obtain more robust features.

3.2 Feature extraction

In the extraction of person image features, ResNet50 is used as the backbone and is denoted as E_v. As shown in Figure 1, the output image features is denoted as $F_v \in^{H\times W\times C}$, and split F_v horizontally into N patches. Where H, W and C denote the length, width, and number of channels in the feature map, respectively. The feature maps of the l-th patch of the i-th Pedestrian image in a batch are represented as $X^l_{v,i} \in R^{H/N\times W\times C}$, where N denotes the total number of patches. We performed maximum pooling on each patch to obtain the feature vector $h^l_{v,i}$.

In the extraction of text features, the pre-trained Bidirectional Encoder Representation from Transformers (BERT) model Kingma and Ba (2014) is used as the backbone. The output text features are $Y_j = (t_{j,0}, t_{j,1}\ldots t_{j,M}) \in R^{(M+1)\times D}$, where $t_{j,1}, \ldots, t_{j,M}$ represents the features of M words. and $t_{j,0}$ represents the global features of Y_j.

3.3 Explicit semantic feature extraction

Since the text describing the same pedestrian may have multiple sentences and inconsistent features after encoding, this unclear semantics leading to ineffective alignment. To address this problem, this paper proposes the Explicit Semantic Feature Extraction (ESFE) module. In general, this module bases on the fact that each divided region of the pedestrian image has clearer semantic information, which we can use to guide the learning of the text features. By aligning the semantic information between text and image features, the proposed module endows text features with clear

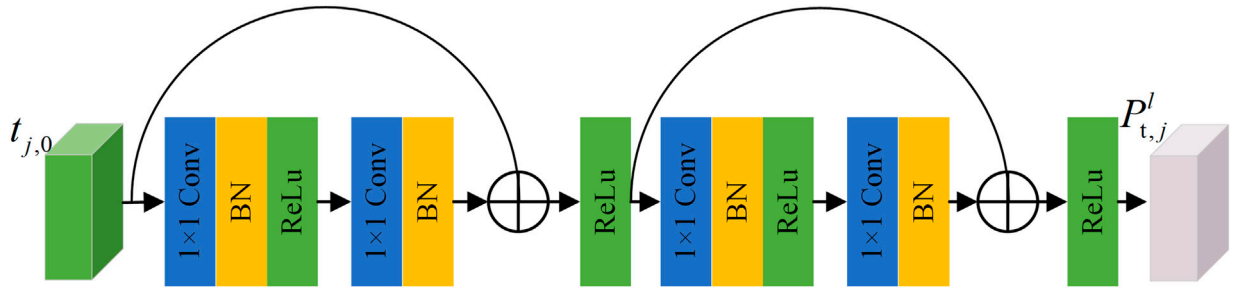


FIGURE 2
Semantic alignment network.

semantic information. Specifically, the Semantic Alignment Network (SAN) is employed in the ESFE module to map classification token to features with local semantics that are consistent with pedestrian image features.

The structure of the semantic alignment network is shown in Figure 2, in order to generate the same amount as the N local features of the image. The Explicit Semantic Feature Extraction (ESFE) module also contains N of Semantic Alignment Network (SAN). Each SAN has its own set of parameters and is used to map the global feature of text to different semantic spaces that correspond to different local features of the pedestrian. Assume that j-th text feature of the pedestrian encoded by the Bert network is $t_{j,0}$, the resulting feature we get from SAN is represent as $P^l_{t,j}$.

It is necessary to ensure that the feature channels of both modalities are the same, and we use 1 × 1convolution to expand the number of channels of $P^l_{t,j}$ to obtain $h^l_{t,j}$. We also use Cross-Modal Projection Matching (CMPM) loss Zhang and Lu [31]. $h^l_{v,i}$ and $h^l_{t,j}$ are matching probabilities which can be calculated by the following equation:

$$S^l_{v2t,i,j} = \frac{\exp\left(\left(h^l_{v,i}\right)^T \bar{h}^l_{t,j}\right)}{\sum_{j=1}^{n} \exp\left(\left(h^l_{v,i}\right)^T \bar{h}^l_{t,j}\right)}, \tag{1}$$

where $\bar{h}^l_{t,i} = \frac{h^l_{t,i}}{\|h^l_{t,i}\|_2}$, and the matching loss from image to text in a mini-batch is computed by:

$$L_{v2t}(E_v) = \frac{1}{n}\sum_{l=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} S^l_{v2t,i,j} \log\left(\frac{S^l_{v2t,i,j}}{z^l_{i,j} + \varepsilon}\right), \tag{2}$$

where $\varepsilon = 10^{-8}$, n is the batchsize,$z^l_{i,j} = y^l_{i,j} / \sum_{j=1}^{N} y^l_{i,j}$, and $y^l_{i,j} = 1$ indicates that both belong to the same ID. The loss in the v2t direction adds the loss in the t2v direction to obtain the CMPM loss. The formula is shown as follows:

$$L^1_{cmpm}(E_v) = L_{v2t}(E_v) + L_{t2v}(E_v), \tag{3}$$

3.4 Information supplementation network learning

To address the issue of information incompleteness in the individual local features, which hinders a comprehensive representation of the features, we propose the Information Supplementation Network (ISN) to enrich the semantic information of the local features and thus enhance the feature representation. For the image modality, the local features $h^l_{v,i}$ with missing semantic information are supplemented using ISN (as shown in Figure 3) to obtain locally complete features $f^l_{v,i}$ which will later be concatenated in a specific channel order to form robust global features $f^g_{v,i}$. We illustrate this process using the k-th visual local feature as an example. First, we compute the similarity between $h^k_{v,i}$ and $h^p_{v,i}\,(p \neq k)$ embedded in a common space:

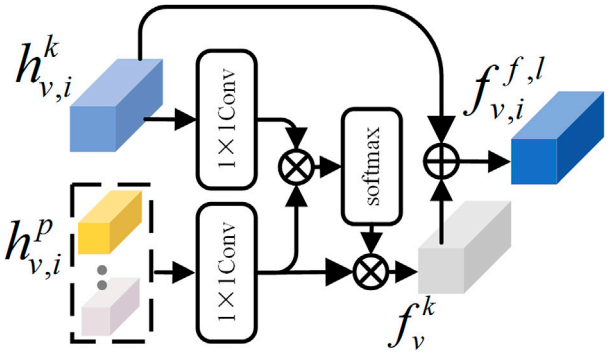


FIGURE 3
Information supplementation network.

$$S_{k,p} = \frac{W_k\left(h^k_{v,i}\right)^T W_p\left(h^p_{v,i}\right)}{\|W_k\left(h^k_{v,i}\right)\|\|W_p\left(h^p_{v,i}\right)\|}, \tag{4}$$

where, W_k, $W_p$are two parameter matrices that can be updated during training. Then, the association strength between the k-th image local feature and the other local features can be expressed as follows:

$$\alpha_{k,p} = \frac{\exp\left(S_{k,p}\right)}{\sum_{p=1,p\neq k}^{N} \exp\left(S_{k,p}\right)}, \tag{5}$$

After extracting the missing information of local feature $h^k_{v,i}$ among N-1 local features using $\alpha_{k,p}$, we can obtain the missing information of the k-th visual local feature f^k_v, with the following equation.

$$f_v^k = W_a\left(\sum_{p=1,p\neq k}^{N}\alpha_{k,p}W_P\left(h_{v,i}^k\right)\right), \tag{6}$$

Finally, the semantic information is refined by fusing the missing information with the original information, and the formula can be expressed as follows.

$$f_{v,i}^k = W_f\left(f_v^k + h_{v,i}^k\right), \tag{7}$$

where W_f, W_a are two learnable matrices. Similar to the visual features, we also process the text local features by the same steps as mentioned above to obtain semantically perfect text local features $f_{t,j}^l$. To ensure the consistency of local features, the following loss function is used to optimize the training.

$$L_{cmpm}^2\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = L_{v2t}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) + L_{t2v}\left(E_v, E_c, W_l, W_p, W_f, W_a\right), \tag{8}$$

$$L_{v2t}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = \frac{1}{n}\sum_{l=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{i,j}^l \log\left(\frac{p_{i,j}^l}{q_{i,j}^l + \varepsilon}\right), \tag{9}$$

$$L_{t2v}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = \frac{1}{n}\sum_{l=1}^{N}\sum_{j=1}^{n}\sum_{i=1}^{n} p_{i,j}^l \log\left(\frac{p_{i,j}^l}{q_{i,j}^l + \varepsilon}\right), \tag{10}$$

$$p_{i,j}^l = \frac{\exp\left(\left(f_{v,i}^l\right)^T f_{t,j}^l\right)}{\sum_{j=1}^{n}\exp\left(\left(f_{v,i}^l\right)^T f_{t,j}^l\right)}, \tag{11}$$

To make the semantics on the global feature channel consistent as well, we concatenate together different local features on the channel in a specific order to form semantically comprehensive global features $f_{v,i}^g$ and $f_{t,j}^g$, and use the following loss function to optimize the network parameters.

$$L_{cmpm}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = L_{v2t}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right) + L_{t2v}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right), \tag{12}$$

where L_{v2t}^g and L_{t2v}^g can be similarly obtained from Eq. 10 Throughout the training process, the final loss function of the model can be expressed as:

$$L\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = L_{cmpm}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right) + \lambda_2 L_{cmpm}^2\left(E_v, E_c, W_l, W_p, W_f, W_a\right) + \lambda_1 L_{cmpm}^1\left(E_v\right), \tag{13}$$

where, λ_1 and λ_2 are used as parameters to balance the importance of different modules.

4 Experiments

4.1 Datasets and evaluation protocols

To verify the effectiveness of our proposed algorithm, we demonstrate its performance on two challenging datasets CUHK-PEDES Li et al. (2017a) and RSTPReid Zhu et al. [32].

CUHK-PEDES: This dataset is the first publicly available dataset for this task. We adopted the same data partitioning strategy as Chen et al. [33], where the dataset was divided into training, validation, and testing sets. The training set contains 11,003 individuals with a total of 34,054 images and 68,126 textual descriptions. Some sample images and text descriptions are shown in Figure 3, Figure 4. The validation set contains 1,000 individuals with 3,078 images and 6,158 textual descriptions, while the testing set contains 1,000 individuals with 3,074 images and 6,156 textual descriptions.

RSTPReid: This dataset is the latest public dataset. This dataset contains 4101 pedestrians with different identities, each with five different images, resulting in a total of 20,505 person images, with two textual descriptions per image. Following the data partitioning strategy in Zhu et al. [32], we divided this dataset into training, validation, and testing sets, where the training set contains 18,505 images from 3,701 individuals, the validation set contains 1,000 images from 200 individuals, and the testing set contains 1,000 images from 200 individuals. Similar to existing methods, we employ the Cumulative Match Characteristic metric to evaluate the performance of our model.

4.2 Implementation details

Our network is mainly composed of image feature extractor and text feature extractor. As with the other methods, we use ResNet-50 trained on imageNet Russakovsky et al. [34] and Bert as the backbone. The network was trained for 100 generations. Optimize network parameters using the Adam optimizer. Kingma and Ba [35]. The initial learning rate is set to 1×10^{-3}, and the warm-up strategy in Luo et al. [36] is used to adjust the learning rate for the first 10 epochs. At the 41st epoch, the learning rate is decayed to 10% of its current value. All images are resized to $384 \times 128 \times 3$, and data augmentation is performed using random horizontal flipping. The batch size is set to 64, with each batch containing 64 image-text pairs. The text length is uniformly set to 64. During testing, cosine distance is used to measure the similarity between image-text pairs. The proposed model is implemented based on the PyTorch. All experiments are conducted on a single NVIDIA GeForce RTX3090 GPU device.

4.3 Comparison with state-of-the-art methods

4.3.1 Results on the CUHK-PEDES dataset

To illustrate the advantages of our method, we perform our method on the CUHK-PEDES dataset, and compare its performance with some state-of-the-art methods. The methods involved in the comparison include GNA-RNN Li et al. [29], GLA Chen et al. [33], CMPM + CMPC Zhang and Lu [31], MCCL Wang et al. [37], A-GANet Liu et al. [12], Dual-path Zheng et al. [38], MIA Niu et al. [19], PMA Jing et al. [39], TIMAM Sarafianos et al. [40], ViTAA Wang et al. [16], NAFS Gao et al. [20], DSSL Zhu et al. [32], MGEL Wang et al. [21], SSAN Ding et al. [10], TBPS(ResNet-50) Han et al. [41], and SUM Wang et al; [42]. The experimental results of different methods are shown in Table 1. It can be observed that the proposed method achieves a Rank-1 accuracy of 61.97 (%) and a

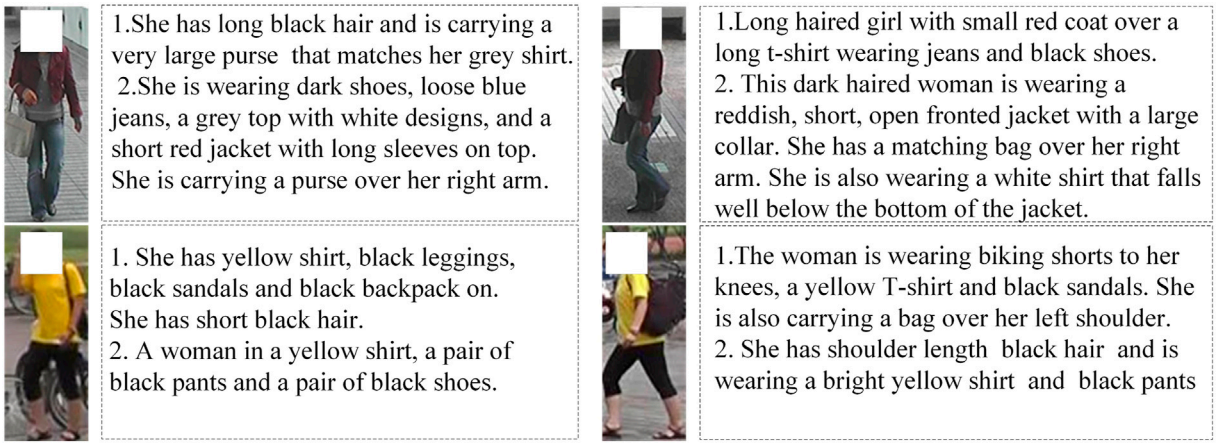


FIGURE 4
Sample CUHK-PEDES dataset display.

Rank-5 accuracy of 81.01 (%), outperforming all the compared methods on the CUHK-PEDES dataset. In addition, it was found that the latest methods NAFS, SSAN and MGEL are far superior to other methods due to the use of attention mechanisms that allow the network to extract robust features adaptively. However, they do not consider the impact of the ambiguous semantic relationship between the textual and imaging descriptions of pedestrians on the matching performance, thus their performance is limited to some extent. Compared with the best-performing method TBPS in the compared methods, the proposed method achieves 0.32 (%) improvement in Rank-1 accuracy, which demonstrates the effectiveness and superiority of the proposed method over the compared methods.

4.3.2 Results on the RSTPReid dataset

In order to further verify the effectiveness of our method, we also conducted a comparative test on the RSTPReid dataset. Our proposed method is compared with five latest methods, namely, IMG-Net Wang et al. [44], AMEN Wang et al. [43], DSSl Zhu et al. [32], SSAN Ding et al. [10], and SUMWang et al.(2022c). As shown in Table 2, the latest method SSAN achieves the best performance with 43.50(%), 67.80(%) and 77.15 (%) accuracy for rank-1, rank-5 and rank-10, respectively. In contrast, the proposed method achieves significantly higher performance with 43.88(%), 76.60(%), and 80.20 (%) accuracy for rank-1, rank-5, and rank-10, respectively, exceeding the performance of SSAN. These experiments further validate the effectiveness of our method.

4.4 Ablation study

The proposed method in this paper mainly consists of two parts: Explicit Semantic Feature Extraction (ESFE), and Information Supplementation Network (ISN). In this paper, we use the model obtained by pre-training ResNet50 and Bert under the constraint of loss function as Baseline, and in order to verify the effectiveness of each module, different modules are added to Baseline gradually to observe the change of matching performance. In this process, the model obtained by adding ESFE to Baseline is named "Baseline + ESFE"; the model obtained by adding ISN to Baseline is named "Baseline + ISN″ The model after adding ISN to "Baseline + ESFE" is "Baseline + ESFE + ISN". All experiments were conducted on the CUHK-PEDES dataset, and the experimental results are shown in Table 3.

4.4.1 The effectiveness of ESFE

In this paper, ESFE is mainly used to address the problem of semantic mismatch between textual features and their corresponding visual objects. As shown in Table 3, without using ESFE, the performance of the Baseline model on rank-1 accuracy is only 55.14 (%). When ESFE is added to the Baseline model, the performance of Baseline + ESFE is improved from 55.14 (%) to 58.42 (%), with an increase of 3.28 (%). This is mainly because ESFE can effectively address the issue of misalignment between features.

4.4.2 The effectiveness of ISN

To supplement local features, the ISN method is proposed in this paper. In this process, various local features are fused by self-attention mechanism to obtain comprehensive features. As shown in Table 3, without using ISN, the performance of the Baseline model on rank-1 accuracy is only 55.14 (%). When ISN is added to the Baseline model, the performance of Baseline + ISN is improved from 55.14 (%) to 59.20 (%), with an increase of 4.06 (%). This is mainly because ISN can effectively supplement missing information in features and improve the comprehensiveness of features.

4.4.3 The effectiveness of ESFE + ISN

Table 3 shows the effectiveness of adding ISN to Baseline + ESFE after rough alignment of local features. It can be seen that supplementing information between roughly aligned local features is more effective than directly supplementing information on the baseline. Rank-1 is improved from 59.20 (%) to 61.97 (%), with an increase of 2.77 (%). This indicates that supplementing information on relatively good features can result in more robust features.

TABLE 1 Comparative experiments on CUHK-PEDES dataset. Where the optimal results are shown in bold.

Methods	References	Rank-1	Rank-5	Rank-10
GNA-RNN	CVPR'17	19.05	—	53.64
Li et al. [29]				
GLA	ECCV'18	43.58	66.93	76.26
Chen et al. [33]				
CMPM + CMPC	ECCV'18	49.27	—	79.27
Zhang and Lu [31]				
MCCL	ICASSP'19	50.58	—	79.06
Wang et al. [37]				
A-GANet	ACM MM'19	53.14	74.03	81.95
Liu et al. [12]				
Dual-path	TOMM'20	44.4	66.26	75.07
Zheng et al. [38]				
MIA	TIP'20	53.10	75.00	82.9
Niu et al. [19]				
PMA	AAAI'20	53.81	73.54	81.23
Jing et al. [39]				
TIMAM	ICCV'20	54.51	77.56	84.78
Sarafianos et al. [40]				
ViTAA	ECCV'20	55.97	75.84	83.52
Wang et al. [16]				
NAFS	arXiv'21	59.94	79.86	86.7
Gao et al. [20]				
DSSL	ACMMM'21	59.98	80.41	87.56
Zhu et al. [32]				
MGEL	IJCAI'21	60.27	80.01	86.74
Wang et al. [21]				
SSAN	arXiv'21	61.37	80.15	86.73
Ding et al. [10]				
TBPS(ResNet-50)	arXiv'21	61.65	80.98	86.78
Han et al. [41]				
SUM	KBS'22	59.22	80.35	87.51
Wang et al. [42]				
Our(Proposed)	This paper	**61.97**	**81.01**	**87.82**

TABLE 2 Comparative experiments on RSTPReid dataset, and the best result is shown in bold.

Methods	References	Rank-1	Rank-5	Rank-10
IMG-Net	JEI'20	37.60	61.15	73.55
Wang et al. [44]				
AMEN	PRCV'21	38.45	62.40	73.80
Wang et al. [43]				
DSSL	ACMMM'21	39.05	62.60	73.95
Zhu et al. [32]				
SSAN	arXiv'21	43.50	67.80	77.15
Ding et al. [10]				
SUM	KBS'22	41.38	67.48	76.48
Wang et al. [42]				
Our(Proposed)	This paper	**43.88**	**76.60**	**80.20**

TABLE 3 Ablation experiment.

Methods	Rank-1	Rank-5	Rank-10
Baseline	55.14	76.64	84.48
Baseline + ESFE	58.42	79.76	85.78
Baseline + ISN	59.20	79.29	85.63
Baseline + ESFE + ISN	61.97	81.01	87.82

4.4.4 Ablation experiments Visualization

Figure 5 presents the effectiveness of each module. It can be observed from Figure 6 that the matching accuracy is improved when ESFE and ISN are added separately to the "Baseline", which demonstrates the effectiveness of the proposed ESFE and ISN. However, the best performance is not achieved, indicating that the current model cannot distinguish finer-grained features. When ISN is added to "Baseline + ESFE", it can be seen that information supplementation on the roughly aligned features can better explore finer-grained features. As shown in Figure 6, the model can not only distinguish large-scale features such as "A short-sleeved red top" and "Short skirt" but also better distinguish finer-grained features such as "Thick heel" and "White logo". This proves the effectiveness of the proposed "Baseline + ESFE + ISN". The above conclusions are consistent with those obtained from Table 3.

4.4.5 Analysis of the loss function

Table 4 shows the effectiveness of the loss function. We find that the Rank-1 of L^1_{cmpm}+L^2_{cmpm} reaches 60.33 (%) and that of L^1_{cmpm}+L^g_{cmpm} reaches 60.12 (%). However, the Rank-1 of L^1_{cmpm} is only 58.42 (%), which we believe is because L^2_{cmpm} and L^g_{cmpm} can train the ISN network better and make the feature information more complete. The Rank-1 of L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm} reaches 61.97 (%) and the Rank-1 of L^1_{cmpm}+L^2_{cmpm} reaches 60.33 (%), which is 1.64 (%) higher, because L^g_{cmpm} constrains the global features and ensures the two modalities consistency of the global features between the two modalities. The Rank-1 of L^1_{cmpm} + L^g_{cmpm} is 60.12 (%), which is lower than the best result. This is because L^2_{cmpm}

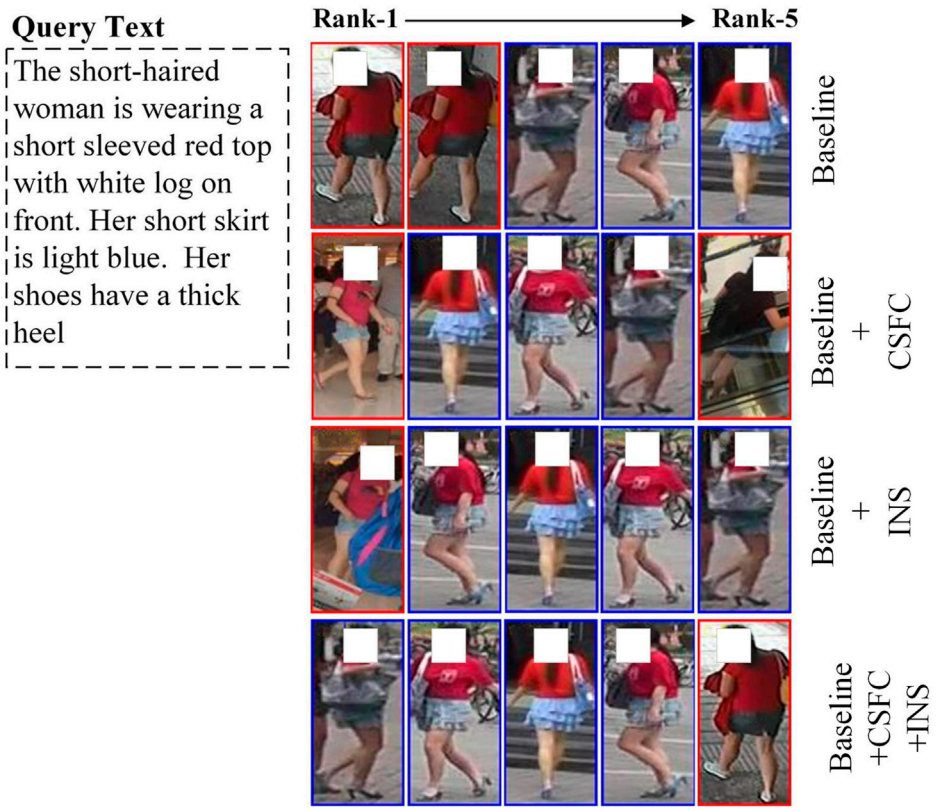


FIGURE 5
Visualize the retrieval results of baseline and our method. The image on the red edge indicates that the query is wrong, and the blue edge indicates that the query is correct.

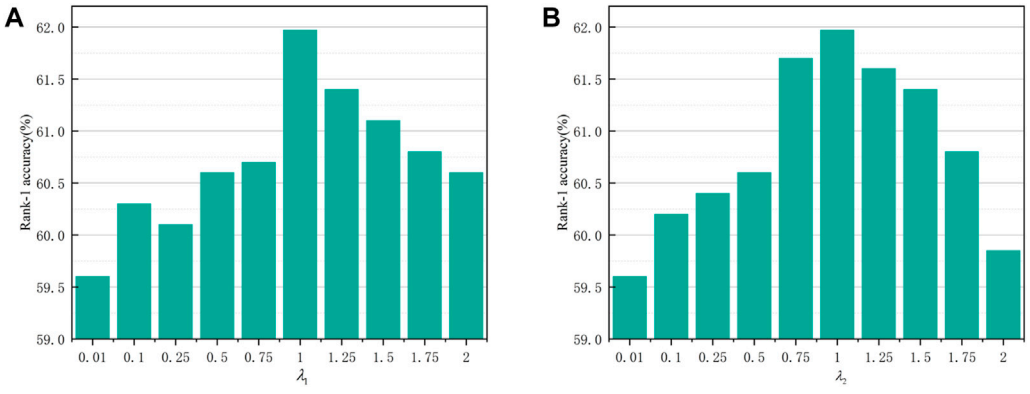


FIGURE 6
Effect analysis on hyperparameters.

constrains each local feature to ensure that each local feature has different semantics within the modality and consistent semantic information between the modalities. Thus $L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm}$ not only ensure that the local features are discriminative, but also ensure that the global features are consistent. This also shows that it is reasonable for us to use $L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm}$ to train the network.

4.5 Parameter selection and analysis

The three main hyperparameters involved in our approach are λ_1, λ_2 and N. In the parametric analysis, we fix two parameters to analyze the effect of another parameter on the results. All our experiments for the parameter analysis were performed on the CUHK-PEDES dataset.

TABLE 4 Analysis of the loss function.

Methods	Rank-1	Rank-5	Rank-10
L^1_{cmpm}	58.42	79.76	85.78
$L^1_{cmpm}+L^2_{cmpm}$	60.33	80.82	86.82
$L^1_{cmpm}+L^g_{cmpm}$	60.12	80.30	86.43
$L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm}$	61.97	81.01	87.82

The influence of λ_1. In Eq. 13, the hyperparameter λ_1 mainly regulates the role played by L^1_{cmpm}. This loss term is used to ensure the initial alignment of each local feature semantics. Figure 6A shows the effect on Rank-1 for different values of λ_1 in the CUHK-PEDES task. From this, we can find that there is an overall improvement in the Rank-1 recognition accuracy of our algorithm on the CUHK-PEDES task when $\lambda_1 \in [0.01, 1]$, the Rank-1 on CUHK-PEDES task decreases when $\lambda_1 \in [1, 2]$ Therefore, $\lambda_1 = 1$ is the optimal choice

The influence of In Eq. 13, the hyperparameter λ_2 mainly regulates the role played by L^2_{cmpm}. This loss term is used to ensure that the ISN can adaptively extract the relationship to each local feature and fuse the features in this way. We fix the hyperparameter $\lambda_1 = 1$, and λ_2 takes values in the range [0,2]. On CUHK-PEDES, the variation of Rank-1 for different values of λ_2 is shown in Figure 6B. It can be seen that when λ_2 is 1, the method in this paper can obtain the optimal performance on CUHK-PEDES, so it is reasonable to set λ_2 to 1.

The influence of *N*. In ESFE, for the image modality, we divide the image features into N local features with different semantics by PCB, and for the text modality, we generate N local features with different semantics by SA network. To verify the effect of different values of *N* on the model performance, we manually set *N* to 2, 3, 4, 6, 8, and12. From which we select the optimal N value for the model performance. This experiment was conducted on the CUHK-PEDES dataset. Table] 5 shows the experimental results of the effect of taking different values on the performance of the model. It can be seen that *N* of 6 achieves the best results.

TABLE 5 The influence of *N*.

N	Rank-1	Rank-5	Rank-10
2	59.14	78.44	86.78
3	60.04	78.89	86.91
4	60.33	79.75	87.44
6	61.97	81.01	87.82
8	61.53	80.93	87.32
10	61.11	79.61	87.14
12	60.47	78.01	86.92

5 Conclusion

This paper proposes a text-based framework for pedestrian image retrieval. Firstly, the ESFE method is utilized to provide clear semantic information for the text and achieve rough alignment between text and image features. In order to further enhance the representation of features, the ISN method is proposed to model the relationships among local features, fuse the features according to the underlying relationships. Finally global features are concatenated by refined local features. This improves the comprehensiveness of the features and effectively alleviates the matching difficulties caused by incomplete features. Compared with existing methods, the proposed model achieves good results on the CUHK-PEDES and RSTPReid datasets. Through ablation study, the contribution of different modules is investigated. The results show that this model is suitable for text-based pedestrian image retrieval. It is worth noting that in our study, sample diversity has a great impact on this task. For future work, we will study how to solve the problem of sample diversity.

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

HZ responsible for paper scheme design, experiment and paper writing. FL guide the paper scheme design and revision. YH guide to do experiments and write papers. XT guide the paper scheme design and revision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Fund Projects: Science and Technology Program of Science and Technology Department of Yunnan Province (Basic Research Special Project) (202101AT070136), National Natural Science Foundation of China (62161015), Key Science and Technology Special Project of Yunnan Province (202002AD080001).

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

References

1. Li H, Xu J, Yu Z, Luo J. Jointly learning commonality and specificity dictionaries for person re-identification. In: IEEE Transactions on Image Processing (2020). p. 7345–58.

2. Zhang L, Li K, Qi Y. Person re-identification with multi-features based on evolutionary algorithm. In: IEEE Transactions on Emerging Topics in Computational Intelligence (2021). p. 509–18.

3. Wang S, Liu R, Li H, Qi G, Yu Z. Occluded person re-identification via defending against attacks from obstacles. IEEE Transactions on Information Forensics and Security (2022).

4. Zhang Y, Wang Y, Li H, Li S. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia (ACMMM) (2022). p. 3347–55.

5. Li H, Yan S, Yu Z, Tao D. Attribute-identity embedding and self-supervised learning for scalable person re-identification. In: IEEE Transactions on Circuits and Systems for Video Technology (2020). p. 3472–85.

6. Zhu Z, Luo Y, Chen S, Qi G, Mazur N, Zhong C, et al. Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification. *J Vis Commun Image Representation* (2021) 80:103303. doi:10.1016/j.jvcir.2021.103303

7. Li H, Kuang Z, Yu Z, Luo J. Structure alignment of attributes and visual features for cross-dataset person re-identification. *Pattern Recognition* (2020) 106:107414. doi:10.1016/j.patcog.2020.107414

8. Lingli L, Minghong X, Fan L, Yafe Z, Huafeng L, Tingting T. *Unsupervised domain adaptive person re-identification guided by low-rank priori*. Chongqing, China: Chongqing University (2021).

9. Li H, Liu M, Hu Z, Nie F, Yu Z. Dupilumab use in non-atopic chronic hand eczema: Two cases and a review of the literature. *IEEE Trans Circuits Syst Video Technol* (2023) 1–3. doi:10.25259/IJDVL_721_2022

10. Ding Z, Ding C, Shao Z, Tao D. *Semantically self-aligned network for text-to-image part-aware person re-identification* (2021). *arXiv preprint arXiv:2107.12666*.

11. Zhang S, Long D, Gao Y, Gao L, Zhang Q, Niu K, et al. *Text-based person search in full images via semantic-driven proposal generation* (2021). *arXiv preprint arXiv:2109.12965*.

12. Liu J, Zha ZJ, Hong R, Wang M, Zhang Y. Deep adversarial graph attention convolution network for text-based person search. In: Proceedings of the 27th ACM International Conference on Multimedia (ACMMM) (2019). p. 665–73.

13. Zheng K, Liu W, Liu J, Zha ZJ, Mei T. Hierarchical gumbel attention network for text-based person search. In: Proceedings of the 28th ACM International Conference on Multimedia (ACMMM) (2020). p. 3441–9.

14. Chen Y, Zhang G, Lu Y, Wang Z, Zheng Y. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing* (2022) 494: 171–81. doi:10.1016/j.neucom.2022.04.081

15. Jing Y, Si C, Wang J, Wang W, Wang L, Tan T. *Pose-guided joint global and attentive local matching network for text-based person search*. New York, NY: AAAI Conference on Artificial Intelligence (AAAI) (2020).

16. Wang Z, Fang Z, Wang J, Yang Y. Vitaa: Visual-textual attributes alignment in person search by natural language. *European conference on computer vision (ECCV)*. Springer) (2020). p. 402–20.

17. Aggarwal S, Radhakrishnan VB, Chakraborty A. Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020). p. 2617–25.

18. Wang C, Luo Z, Lin Y, Li S. Improving embedding learning by virtual attribute decoupling for text-based person search. *Neural Comput Appl* (2022) 34:5625–47. doi:10.1007/s00521-021-06734-9

19. Niu K, Huang Y, Ouyang W, Wang L. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Trans Image Process* (2020) 29:5542–56. doi:10.1109/tip.2020.2984883

20. Gao C, Cai G, Jiang X, Zheng F, Zhang J, Gong Y, et al. *Contextual non-local alignment over full-scale representation for text-based person search* (2021). *arXiv preprint arXiv:2101*.03036.

21. Wang C, Luo Z, Lin Y, Li S. Text-based person search via multi-granularity embedding learning. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI) (2021). p. 1068–74.

22. Li S, Cao M, Zhang M. *ICASSP 2022-2022 IEEE international conference on acoustics*. Speech and Signal Processing (ICASSP) (IEEE) (2022). p. 2724–8.Learning semantic-aligned feature representation for text-based person search.

23. Li H, Chen Y, Tao D, Yu Z, Qi G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Trans Inf Forensics Security* (2021) 16:1480–94. doi:10.1109/tifs.2020.3036800

24. Liu X, Cheung YM, Hu Z, He Y, Zhong B. Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval. *IEEE Trans Emerging Top Comput Intelligence* (2020) 5:607–19. doi:10.1109/tetci.2020.3007143

25. Tang H, Li Z, Peng Z, Tang J. Blockmix: Meta regularization and self-calibrated inference for metric-based meta-learning. In: Proceedings of the 28th ACM international conference on multimedia (2020). p. 610–8.

26. Tang H, Yuan C, Li Z, Tang J. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition* (2022) 130:108792. doi:10.1016/j.patcog.2022.108792

27. Zha Z, Tang H, Sun Y, Tang J. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. In: IEEE Transactions on Circuits and Systems for Video Technology (2023).

28. Li Z, Tang H, Peng Z, Qi GJ, Tang J. Knowledge-guided semantic transfer network for few-shot image recognition. In: IEEE Transactions on Neural Networks and Learning Systems (2023).

29. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X. Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). p. 1970–9.

30. Li S, Xiao T, Li H, Yang W, Wang X. Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017). p. 1890–9.

31. Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018). p. 686–701.

32. Zhu A, Wang Z, Li Y, Wan X, Jin J, Wang T, et al. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia (ACMMM) (2021). p. 209–17.

33. Chen D, Li H, Liu X, Shen Y, Shao J, Yuan Z, et al. Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018). p. 54–70.

34. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* (2015) 115:211–52. doi:10.1007/s11263-015-0816-y

35. Kingma DP, Ba J. *Adam: A method for stochastic optimization* (2014). *arXiv preprint arXiv:1412.6980*.

36. Luo H, Gu Y, Liao X, Lai S, Jiang W. Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019). p. 0.

37. Wang Y, Bo C, Wang D, Wang S, Qi Y, Lu H. Language person search with mutually connected classification loss. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE) (2019). p. 2057–61.

38. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen YD. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans Multimedia Comput Commun Appl (Tomm)* (2020) 16:1–23. doi:10.1145/3383184

39. Jing Y, Si C, Wang J, Wang W, Wang L, Tan T. Pose-guided multi-granularity attention network for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 34 (2020). p. 11189–96. doi:10.1609/aaai.v34i07.6777

40. Sarafianos N, Xu X, Kakadiaris IA. Adversarial representation learning for text-to-image matching. In: Proceedings of The IEEE/CVF International Conference on Computer Vision (ICCV) (2019). p. 5814–24.

41. Han X, He S, Zhang L, Xiang T. *Text-based person search with limited data* (2021). *arXiv preprint arXiv:2110.10807*.

42. Wang Z, Zhu A, Xue J, Jiang D, Liu C, Li Y, et al. Sum: Serialized updating and matching for text-based person retrieval. *Knowledge-Based Syst* (2022) 248:108891. doi:10.1016/j.knosys.2022.108891

43. Wang Z, Xue J, Zhu A, Li Y, Zhang M, Zhong C. Amen: Adversarial multi-space embedding network for text-based person re-identification. In: *Chinese conference on pattern recognition and computer vision (PRCV)*. Springer) (2021). p. 462–73.

44. Wang Z, Zhu A, Zheng Z, Jin J, Xue Z, Hua G. Img-net: Inner-cross-modal attentional multigranular network for descriptionbased person re-identification. *J Electron Imaging* (2020) 29:043028. doi:10.1117/1.jei.29.4.043028

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 30 May 2023
DOI 10.3389/fphy.2023.1208781

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Jian Sun,
Southwest University, China
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China
Yong Li,
Chongqing University, China

*CORRESPONDENCE
Yong Xu,
yongxu@ymail.com

RECEIVED 19 April 2023
ACCEPTED 09 May 2023
PUBLISHED 30 May 2023

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

Multi-level semantic information guided image generation for few-shot steel surface defect classification

Liang Hao[1], Pei Shen[2], Zhiwei Pan[2] and Yong Xu[1]*

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, [2]HBIS Digital Technology Co., Ltd., Shijiazhuang, China

Surface defect classification is one of key points in the field of steel manufacturing. It remains challenging primarily due to the rare occurrence of defect samples and the similarity between different defects. In this paper, a multi-level semantic method based on residual adversarial learning with Wasserstein divergence is proposed to realize sample augmentation and automatic classification of various defects simultaneously. Firstly, the residual module is introduced into model structure of adversarial learning to optimize the network structure and effectively improve the quality of samples generated by model. By substituting original classification layer with multiple convolution layers in the network framework, the feature extraction capability of model is further strengthened, enhancing the classification performance of model. Secondly, in order to better capture different semantic information, we design a multi-level semantic extractor to extract rich and diverse semantic features from real-world images to efficiently guide sample generation. In addition, the Wasserstein divergence is introduced into the loss function to effectively solve the problem of unstable network training. Finally, high-quality defect samples can be generated through adversarial learning, effectively expanding the limited training samples for defect classification. The experimental results substantiate that our proposed method can not only generate high-quality defect samples, but also accurately achieve the classification of defect detection samples.

KEYWORDS

few-shot steel surface defect classification, adversarial learning, residual module, multi-level semantic feature extractor, Wasserstein divergence

1 Introduction

Steel is an essential material for industrial production, with a broad range of uses in areas such as automobile, aerospace and machinery. As the demand for material fitness in various industries increases, the surface quality of steel has become increasingly important. However, during the steel manufacturing process, due to the influence of various unstable factors such as raw materials and production conditions, various types of defects may appear on the surface of steel, which affect the quality of steel to varying degrees and easily lead to serious production accidents, resulting in immeasurable losses to producer and users [1, 2]. Thus, it is of great importance to classify the defects on the surface of steel efficiently for further quality enhancement.

Generally, steel surface defects belonging to the same category meet a large intra-class difference, while those of different categories are highly similar [3], making the classification

of steel surface defects more complicated. To address this problem, various approaches have been studies. For instance, Zaghdoudi et al. [4] proposed a steel surface defect classification method based on the binary Gabor pattern (BGP) algorithm and support vector machine (SVM). Hu et al. [5] extracted various visual features such as geometry, texture, and shape of the defect image and fed them to SVM for classification. Despite the fact that these methods do classify different defects, these hand-crafted features are not optimal, making a constraint on the further performance improvement. Fortunately, thanks to the development of deep learning, deep learning based methods have attracted much attention in the field of steel surface defect classification due to it powerful capability in feature extraction. Specifically, Duan et al. [6] used RGB images and gradient images as inputs to a dual-flow convolutional neural network, and fused multi-source information to recognize aluminum surface defects. Liu et al. [7] proposed an improved dual CNN model fusion framework, which uses pre-trained VGG16 and AlexNet to extract different features from the input source to classify and identify aluminum surface defects.

Although deep learning based methods enjoy superiority compared with conventional methods, they also meet the limitation on the large scale of training data. However, the number of non-defective samples in actual industrial production environments is far greater than that of defective samples. Moreover, it is difficult to identify and collect defective samples, further leading to an insufficient number of samples [8, 9, 10]. To address this issue of insufficient samples, many researchers have begun to focus on the unsupervised data enhancement algorithm: Generative Adversarial Networks (GANs). Currently, many improved GANs and adversarial learning strategies have been derived, such as Wasserstein GAN (WGAN) [11], Deep Convolutional GAN (DCGAN) [12], and ACGAN [13]. These generative models augment the original data by generating synthetic samples, thereby mitigating the effect of few-shot on the classification performance and improving the accuracy. Dosovitskiy et al. [14] showed that even with low-fidelity images, the performance can be significantly improved. If the generated images enjoy the high-quality, the over-fitting problem can further be solved [15]. However, despite the wide application of GANs and its related improved models, there are still some tough difficulties, such as insufficient model feature capture capability, gradient disappearance, and model collapse, etc.

Furthermore, generating high-quality data similar to the original data distribution can solve the over-fitting problem, and enhance the detection accuracy and generalization ability of the model [15]. Lu and Su [16] proposed a novel method to eliminate mura patterns from defect images by using conditional generation adversarial networks; Li et al. [17] studied a cross-domain fault diagnosis method based on deep neural networks, which has a good industrial application prospect; Liu et al. [18] introduced an attention mechanism into feature extraction, proposed a structural defect detection framework based on GAN-CNN, and achieved satisfactory results. Despite the wide application of GANs and its related improved models, there are still some tough difficulties, such as insufficient model feature capture capability, gradient disappearance, and model collapse, etc.

Aimed at above problems, we propose a steel surface defect classification method based on residual adversarial learning with Wasserstein divergence. First, the residual module is introduced into the network framework of adversarial learning, to enhance the feature extraction ability of the model and improve the quality of generated samples. Subsequently, to extract semantic information from defect samples at different levels, we design a multi-level semantic feature extractor (MSFE), which guides sample generation by extracting the most relevant semantic features from images. Then Wasserstein divergence is used to alleviate gradient disappearance, gradient explosion and mode collapse during model training. Finally, high-quality samples are generated, and few-shot steel surface defect classification is realized by adversarial learning. The experimental results show that the proposed method improves the accuracy of steel surface defect classification, which are superior to many state-of-the-arts.

The main contributions of this paper are as follows:

- The residual module is introduced into the network structure of adversarial learning to contribute to the feature extraction. Moreover, multiple convolutional layers are employed in the model architecture to replace the original classification layer, further boosting the classification performance of the model.
- A multi-level semantic feature extractor (MSFE) which effectively extracts features at different levels is designed, fully capturing diverse semantic information of images to guide the generator in sample generation and improve the quality of generated samples.
- The proposed method can generate high-quality samples to compensate for the deficiencies under few-shot conditions, further improving the classification performance.

2 Related works and preliminary knowledge

2.1 Steel surface defect classification

Steel surface defect classification based on deep learning has gained considerable attention in recent years and achieved remarkable results. Chenon et al. [19] proposed a defect classification approach based on a single convolutional neural network, which can extract effective features for defect classification without the prior of hand-crafted features. Nakazawa et al. [20] proposed a method for surface defect classification and image retrieval using convolutional neural networks. The model was trained, validated, and tested using generated data samples, and it was demonstrated that the model trained by synthetic data can be classified efficiently. Zhu et al. [21] studied an intelligent identification algorithm based on convolutional neural networks and random forest algorithms, which enabled the intelligent identification of weld surface defects. However, obtaining effective defect samples is very challenging in the actual industrial environment, and there is the problem of insufficient samples, which leads to a low performance of the surface defect classification model based on deep learning. Therefore, data augmentation and transfer learning have been proposed by many researchers to address the few-shot problem in this field. Wan et al. [22] studied an improved VGG19 neural network based on small samples and unbalanced datasets for strip

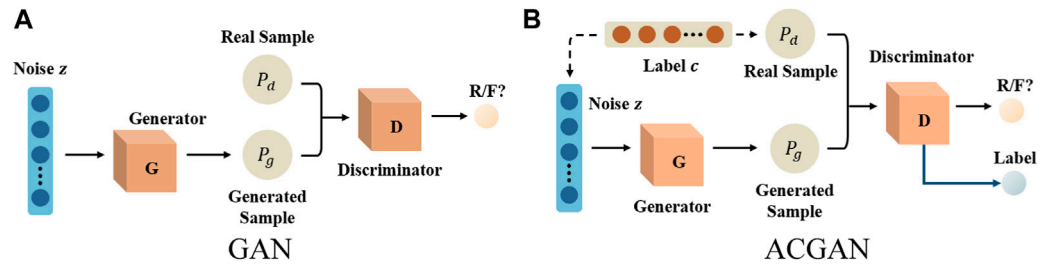


FIGURE 1
The model framework of GAN and ACGAN. **(A)** GAN. **(B)** ACGAN.

steel defect detection. Through fast image preprocessing algorithms and transfer learning theory, excellent results have been achieved on multiple datasets. Han et al. [23] proposed a new framework for intelligent fault diagnosis, namely, Deep Transfer Network (DTN), which generalized deep learning models to domain self-adaptation scenarios. By using the discriminative structure associated with the labeled data in the source domain to adapt to the unlabeled data, more accurate distribution matching is ensured. Furthermore, Liu et al. [24] designed ImDeep, a deep learning model for unbalanced multi-label surface defect classification, which combines three key technologies to improve the classification performance of the model: imbalanced sampler, Fussy-FusionNet, and transfer learning.

Apart from the domain adaptation, some scholars also utilize GAN as a data augmentation technique to address few-shot issue. Goodfellow et al. [15] first proposed the unsupervised deep learning model GAN in 2014, which was inspired by the two-player zero-sum game in game theory and consists of two components: the generator and the discriminator. The generator is mainly responsible for generating data that is as similar as possible to the original data samples, while the discriminator is tasked with distinguishing between real and fake images. Currently, GAN has been widely applied in various fields, such as image generation, data augmentation, image restoration, and image coloring. Specifically, Jain et al. [25] trained three GAN architectures to generate synthetic images for data augmentation, which significantly improved the performance of surface defect classification. He et al. [26] proposed a semi-supervised learning for defect classification based on GAN and ResNet to expand the training samples and exploit the unlabeled images. Zhao et al. [27] designed a reconstruction network to reconstruct the potential defect areas in the sample image, and determine the final defect area according to the difference between the reconstructed sample and the original sample. Lian et al. [28] proposed a novel machine vision method for automatic identification of tiny defects in a single image. To effectively achieve pixel-level defect detection on textured surfaces without manual annotation, Tsai et al. [29] introduced a two-stage deep learning scheme. Particularly, the first stage used CycleGAN to automatically synthesize and annotate the pixels of defect in images. The second stage used the synthesized defect images and their corresponding annotation results as input-output pairs for training the U-Net semantic network.

2.2 Preliminary knowledge

GAN consists of a generator and a discriminator [15], as shown in Figure 1A. The input of the generator is a random noise vector z, and the output is the fake sample generated by it. The discriminator uses the fake sample generated by the generator and the real data x as the input, and the output is the discrimination score of the discriminator on the fake sample. GAN's overall objective function is:

$$\min_G \max_D L(G,D) = E_{x\sim P_d}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z)))] \tag{1}$$

where P_d is the probability density distribution of the real data x; z is the noise vector randomly sampled from the prior distribution P_z; G represents the generator, D represents the discriminator, and $E(\cdot)$ represents the calculated expected value; $D(X)$ is a probability distribution, that is, the probability of classifying data X as a real sample, and X is derived from a real sample x or a generated sample $G(z)$.

Formula 1 shows that the optimization problem of GAN is same as the max-min optimization problem, which includes the optimization goals of the generator and the discriminator. The main function of the discriminator is to perform binary classification on the input data to determine whether the input data comes from the distribution of the real data or the generated pseudo data. Thus, its objective function is:

$$\max_D L(G,D) = E_{x\sim P_d}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z)))] \tag{2}$$

It can be seen from Formula 2 that the goal of the discriminator is to maximize the discrimination accuracy for the data. In other words, we aim to maximize the discriminant result $D(x)$ for the real data x, and minimize the result $D(G(z))$ of the generated sample $G(z)$ (maximize $1-D(G(z))$).

The purpose of the generator is to generate samples that the discriminator cannot distinguish as false, and its objective function is:

$$\min_G L(G,D) = E_{z\sim P_z}[\log(1-D(G(z)))] \tag{3}$$

The generator is optimized by Eq. 3. Specifically, the probability score $D(G(z))$ of the discriminator for the generated sample $G(z)$ is maximized ($1-D(G(z))$ is minimized). During training, the

alternate optimization methods are used: fix one side and update the parameters of the other network. In other words, the model updates the discriminator's parameters firstly through the fixed generator so that the discriminator maximizes the discriminant result. Then we fix discriminator's parameters for updating the generator, which minimize the result that discriminator works. Finally, when the probability distribution P_g of the samples generated by the generator G is infinitely close to the probability distribution P_d of the real samples (that is, $P_g = P_d$), the global optimal solution can be reached.

ACGAN is a variant of GAN [13], and its structure is illustrated in Figure 1B. By incorporating auxiliary label information c into the generator, the generated samples can be constrained to possess certain characteristics, thus allowing for more precise expression of the samples and the generation of specific samples according to it. Moreover, in order to ensure accurate classification, ACGAN adds a softmax layer to the discriminator network, thus enabling the improved model to not only judge the authenticity of the data, but also classify the input samples.

The loss function of ACGAN consists of two parts: the discriminative loss L_s and the classification loss L_c. The role of discriminative loss is to judge the authenticity of the generated samples, thereby improving the quality of the samples generated by the generator. The role of the classification loss is to measure the accuracy of the classification of the sample category. And, the specific calculation of L_c is:

$$L_c = E_{x\sim P_d}[R(x|c_x)] + E_{z\sim P_z,\ c\sim P_c}[R(G(z,c)|c)] \tag{4}$$

where R is the cross-entropy loss function, c_x represents the category label of the real data x, c is the category label of the generated data $G(z,c)$, and P_c is the category label distribution of the sample.

Since a classifier is added to the discriminator D, the network can not only distinguish the authenticity of the data, but also classify the data, so its loss function needs to calculate two parts: discriminant loss $L_s(D)$ and classification loss L_c. The specific calculation is as follows.

$$L_s(D) = E_{x\sim P_d}[\log D(x)] + E_{z\sim P_z,\ c\sim P_c}[\log(1 - D(G(z,c)))] \tag{5}$$

$$L(D) = L_c + L_s(D) \tag{6}$$

Similarly, the loss function of the generator G also needs to consider the classification loss:

$$L_s(G) = E_{z\sim P_z,\ c\sim P_c}[\log(1 - D(G(z,c)))] \tag{7}$$

$$L(G) = L_c - L_s(G) \tag{8}$$

Formulas 6, 8 ultimately constitute the entire loss function of the ACGAN model. During the training process, the model is continually optimized to enhance the quality of the samples generated by the model and augment the classification accuracy of the model.

3 Methods

Although Generative Adversarial Networks (GANs) and Auxiliary Classifier GANs (ACGANs) can effectively alleviate the few-shot classification problem by generating samples, they still meet the limitation on inadequate information extraction capabilities, gradient vanishing, and pattern collapse. To address these issues, we propose a novel network structure. Specifically, a residual adversarial learning model with Wasserstein divergence based on ACGAN under multi-level semantic guidance is proposed, as shown in Figure 2.

First, the random noise vector z and sample label c are input into the generator. The generator generates synthetic samples I_g, expanding the scale of the training data. By utilizing a multi-level semantic feature extractor to process original samples, semantic and contextual information can effectively be captured and used for guiding sample generation of generator. Then, the discriminator takes the generated sample I_g and real sample I as the inputs, and outputs the discriminant result $R/F?$ (True or Fake) and the classification result c' of the generated sample. During the adversarial training of model, the Wasserstein divergence (W_div) is used as the distance measurement between the distributions of the initial data and the distributions of the generated data.

3.1 The modification of network

Despite the fact that ACGAN achieved significantly satisfactory results in image generation [13], it still faces the problem of insufficient feature extraction ability when it is applied to tasks within the few-shot environment, resulting in inadequate acquisition of image information and a consequent decrease in model performance. To address this issue, the overall network structure of ACGAN is optimized, as illustrated in Figure 3. The specific improvements of the network structure are detailed below.

(1) As shown in Figure 3, the residual module (Residual) is introduced into the network structure of the generator and the discriminator to optimize the feature learning ability of the model, so that the model can extract more valuable features. Meanwhile, it can ensure the quality of the samples generated by the model while optimizing the model's ability to discriminate and classify images. The specific network structure of the introduced residual module is shown in Figure 4.

(2) When the kernel size of the deconvolution layer cannot be divisible by stride in the actual calculation, uneven overlapping problems will occur. Also, the generated sample images would have some checkerboard-like artifacts [30]. Therefore, in order to avoid such problems, as shown in Figure 3A, the up-sampling layer (US) and the convolutional layer (Conv) are used to generate sample images in the generator network structure. As shown in Figure 3B, in the discriminator network structure, two convolutional layers are added before the sigmoid and softmax classification layers, which makes the classifier in the discriminator learn more image information and improve the classification performance.

(3) The generator network mainly consists of several residual modules and convolutional layers as well as operating up-sampling layers. The input of the model is the randomly generated 128-dimensional vector z and the sample label c, which undergoes a fully connected layer (FC) and the reshape (reshape) operation. Before the convolution calculation, the first two convolution layers have

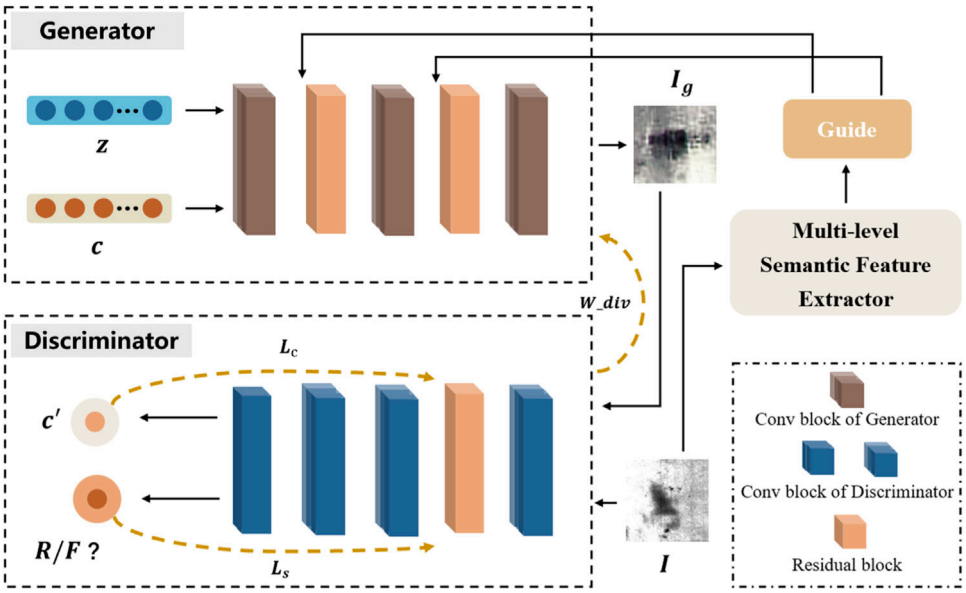


FIGURE 2
Overview of our framework. Given an image *I* as input, our framework first extracts rich semantic information through multi-level semantic feature extractor to guide generator. After that, we deliver the noise *z* and label *c* to generator for generating sample I_g. Finally, we can obtain the classification result *c′* and the discriminate result *R/F*? (True or Fake?) of generated sample I_g by discriminator. L_c, L_s, and *W_div* indicates respectively the classification loss, the discriminant loss, Wasserstein divergence during training.

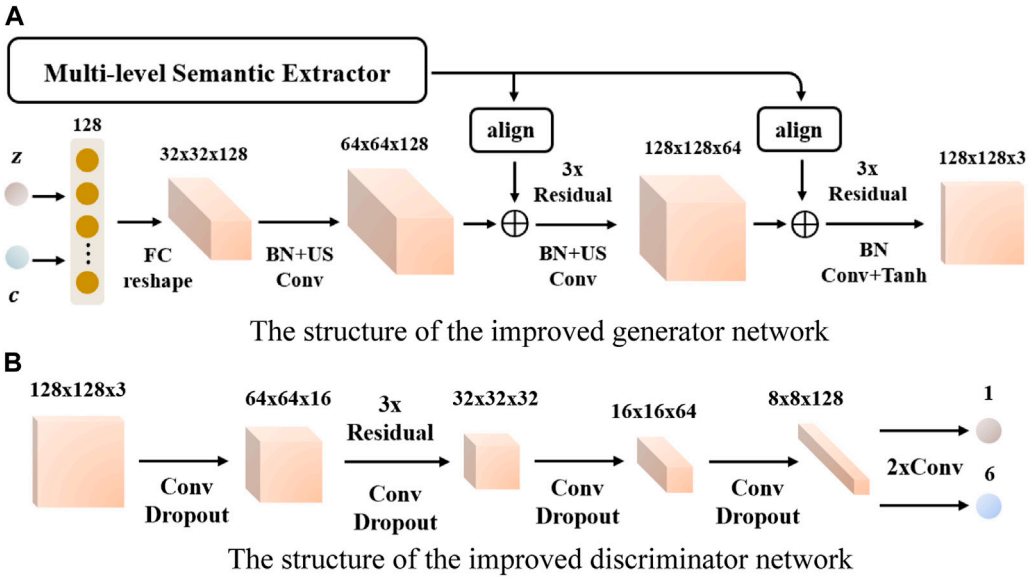


FIGURE 3
Improved generator and discriminator network structure. **(A)** The structure of the improved generator network. **(B)** The structure of the improved discriminator network.

both performed the up-sampling operation using the nearest neighbor interpolation, which increases the feature map by two times. At the same time, Batch Normalization (BN) is used to optimize the network throughput the training. There are three residual modules between each two convolution layers to improve the feature learning ability of the model, and the Leaky-ReLU activation function is used between each layer. The discriminator network also includes 6 convolutional layers and 3 residual modules. And the 3 residual modules follow the first convolutional layer. At the same time, a Dropout layer (Dropout) is further introduced to prevent overfitting problems. Furthermore, the Leaky-ReLU activation function is used between each layer.

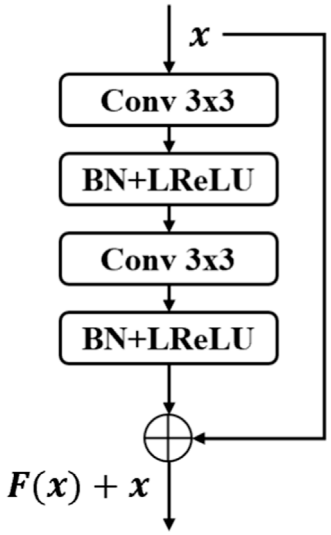


FIGURE 4
The specific network structure diagram of the residual block.

3.2 Multi-level semantic feature extractor

Recently, ACGAN has achieved remarkable progress in the field of image generation. Given a category label, ACGAN can map random noise into high-resolution images with abundant texture features and comprehensive shape details. However, satisfactory results depend on training ACGAN with sufficient quantity of samples. When there is an inadequate number of samples, the effectiveness of ACGAN in generating samples close to reality is compromised due to its inability to obtain enough semantic information, which motivates us to design a multi-level semantic feature extractor to facilitate sample generation tasks, as shown in Figure 2. As illustrated above, the role of the multi-level semantic feature extractor is to extract the semantic and contextual information of defects at different levels in the image. Therefore, the original samples are input into the multi-level semantic feature extractor to obtain the learned hierarchical features, such as texture and shape, which are then incorporated into the generator to serve as guidance for sample generation. Specifically, the sample image I corresponding to the true sample label c is processed by the multi-level semantic feature extractor to obtain rich semantic information, which is then aligned with the different convolutional layers in the generator, facilitating better integration of multi-level semantic features into the process of sample generation. The core of alignment operation mainly relies on a convolutional layer, which adjusts semantic features extracted by MSFE to the size corresponding to different layers of the generator, and adds the adjusted features to the original ones to obtain new features under semantic guidance. The aligned features are added to the convolutional layers of the generator, leveraging diverse levels of semantic features to facilitate the generation of specific defective samples, as depicted in Figure 3A. We use VGG19 pretrained on the ImageNet dataset as a multi-level semantic feature extractor, and use the features extracted from layers 7 to 23 in it to guide the generator.

3.3 Objective function

The Kullback-Leibler (KL) divergence [15] is prone to gradient instability in the Generative Adversarial Networks (GANs) training phase, and can also lead to mode collapse. To address these issues, the Wasserstein GAN (WGAN) uses the Wasserstein distance to ensure that the gradient of the model is continuous during the training process [11]. However, WGAN utilizes weight clipping to restrict the weights within a fixed range strictly, which greatly limits the expressiveness of the network. Consequently, WGAN-GP [31] adopts gradient penalty to enhance the stability of the network training. According to the research conducted by [32], in experiments, WGAN-GP typically employs the technique of interpolating between real and fake data to simulate a uniform distribution across the whole space. This approach is somewhat mechanistic and empirical, which makes it challenging to simulate the full spatial distribution using limited sampling.

In order to solve this problem, Wu et al. [32] proposed Wasserstein divergence to reduce the distance loss function properly between two distributions, as shown in Formula 9. It removes the K-Lipschitz conditional restriction, and changes the penalty term added to the loss function.

$$W_{k,p}(P_d, P_z) = \max_D E_{x\sim P_d}[D(x)] - E_{z\sim P_z}[D(z)] - kE_{u\sim P_u}\left[\|\nabla D(u)\|^p\right] \tag{9}$$

where k and p are selected empirically. Generally, k = 2, p = 6. ∇ represents the gradient. x comes from the distribution P_d of the real data; similarly, z comes from the generated sample distribution P_z. P_u is a distribution derived from the real data distribution P_d and the generated data distribution P_z. D represents the discriminator, and $E(\cdot)$ represents the calculated expected value. Experiments in [32] prove that all different distributions have improved performance.

Based on the loss function of ACGAN [13], we use Wasserstein divergence to address the potential gradient explosion issue in the training process. Hence, the loss function of our method consists of two parts: the loss function $L(D)$ of discriminator and the loss function $L(G)$ of generator, with each loss function consisting of two components: the adversarial loss function L_s and the conditional loss function L_c.

The purpose of $L(D)$ is to ensure that the discriminator can distinguish between real and generated samples and accurately classify them based on their respective conditions, as shown below:

$$L_s(D) = E_{x\sim P_d}[D(x)] - E_{z\sim P_z}[D(z)] - kE_{u\sim P_u}\left[\|\nabla D(u)\|^p\right] \tag{10}$$

$$L_c = E_{x\sim P_d}[R(x|c_x)] + E_{z\sim P_z,\, c\sim P_c}[R(G(z,c)|c)] \tag{11}$$

$$L(D) = L_c + L_s(D) \tag{12}$$

where $L_s(D)$ represents the adversarial loss function that is modified with Wasserstein divergence; L_c is the conditional loss function; $R(\cdot)$ denotes the cross-entropy loss function; c_x indicates the category label of real data sample x, and c denotes the category label of generated data $G(z,c)$. P_c represents the distribution of sample class labels. During the training process of discriminator, our objective is to maximize its loss function $L(D)$.

Likewise, the purpose of $L(G)$ is to generate high-quality data samples such that the discriminator cannot distinguish whether the sample is real or fake, as illustrated below:

$$L_s(G) = E_{z\sim P_z}[D(z)] \tag{13}$$

$$L_c = E_{x\sim P_d}[R(x|c_x)] + E_{z\sim P_z,\ c\sim P_c}[R(G(z,c)|c)] \tag{14}$$

$$L(G) = L_c - L_s(G) \tag{15}$$

where $L_s(G)$ represents the adversarial loss function for the generator. Similarly, we aim to maximize its loss function $L(G)$ in the training process.

3.4 Network training

During the training process of the model, the discriminator continuously enhances its capability to distinguish between real samples and generated samples, while the generator continuously improves its ability to generate realistic samples. The discriminator updates its weights by utilizing both real and generated samples, and the generator updates its weights through the error feedback from the discriminator. The training process of the model is a maximization and minimization process. In the adversarial training of the discriminator and the generator, the discriminator minimizes the probability of misclassification, and the generator maximizes the error probability of the discriminator. The iterative training method of the generator and the discriminator is employed to prevent the over-fitting of the generator network. The specific training steps of the model are illustrated in Algorithm 1.

Data: image dataset
Output: trained Discriminator and Generator, Training Accuracy

for *epoch=0* **to** *n* **do**
  randomly sample from real samples and get (*real_images*, *labels*), and randomly sample from a uniform distribution to obtain noise *z*
  input (*z*, *labels*) into Generator to generate sample *fake_images*
  generated sample *fake_images* and real sample *real_images* are fed into discriminator
    calculate the gradient of the real sample space, calculate the gradient of the generated sample space, and calculate the Wasserstein divergence according to Formula 9
  for *D_epoch=0* **to** *m* **do**
    calculate Discriminator's loss by Formulas 10, 11, 12
    update Discriminator parameters
  end for
    calculate Generator's loss according to Formulas 13, 14, 15
    update Generator parameters
  end for

Algorithm 1. Residual Adversarial Learning Model with Wasserstein Divergence.

4 Experiments

In order to verify the effectiveness of the proposed method, experiments are conducted on the NEU-CLS dataset using a Windows 10 system with 16 GB of memory, an AMD Ryzen 7 4800HS processor, and an NVIDIA GTX 1660 Ti graphics card. The model is constructed using the PyTorch platform.

4.1 Dataset

This paper performs experiments on the NEU-CLS hot-rolled steel surface defect dataset from Northeastern University [34]. The dataset consists of 6 types of defects, and each category contains 300 grayscale images (200 × 200 pixels). These six types of defects are: crazing (Cr), inclusion (In), patches (Pa), pitted surface (PS), rolled-in scale (RS) and scratches (Sc), as illustrated in Figure 5.

In the experiment, the NEU-CLS dataset is divided according to a 2:1 ratio, with 1,200 images used as the training set and 600 images used as the test set. It takes 10,000 epochs to train our network with Adam optimizer and a batch of 64 images. The parameter settings of the model are as follows: learning rate of $\alpha = 0.0002$, random noise vector dimension of z = 128, and Adam optimization parameters of $\beta1 = 0.5$ and $\beta2 = 0.999$. In addition, we use VGG19 pretrained on the ImageNet dataset as a multi-level semantic feature extractor, and use the features extracted from layers 7 to 23 in it to guide the generator.

4.2 Few-shot classification of steel surface defects

Considering the restricted size of the dataset, we conduct experiments with six different training sample sets (200, 150, 100, 50, 30, 10) to evaluate the few-shot classification performance enhancement of the proposed method after training, and to comparatively analyze the impact of the data size on the model. The numbers of test sets are kept constant. The results of the comparison between ACGAN and the method proposed in this paper under different training sample sizes are presented in Table 1.

According to Table 1, it can be observed that the classification performance of ACGAN and the proposed model decreases as the training sample size decreases. It is evident that insufficient samples reduce the generalization capability of the model, resulting in a poorer performance on the test set. Furthermore, the decline of our model is more gradual than that of ACGAN, indicating that the method proposed in this paper is more stable and robust when dealing with few-shot issues. As illustrated in Figure 6, the trend of classification results of ACGAN and our model under different training sample sizes can be observed.

Observing Figures 5, 6, it can be seen that the accuracy of our model has a distinct advantage over ACGAN under different training sample sizes. When the sample size is 200, the average accuracy of our model reaches 98.67%. At the same time, when the training sample size is 10, the average accuracy of the model in this paper is 89.67%, while the accuracy of ACGAN drops to 66.5%. This indicates that ACGAN is more reliant on data. Furthermore, as the training sample size decreases, the classification accuracy gap between ACGAN and the model proposed in this paper increases. When the sample size is 10, the accuracy of ACGAN is 23.17% lower than that of the method proposed in this paper, making it evident that ACGAN is far less

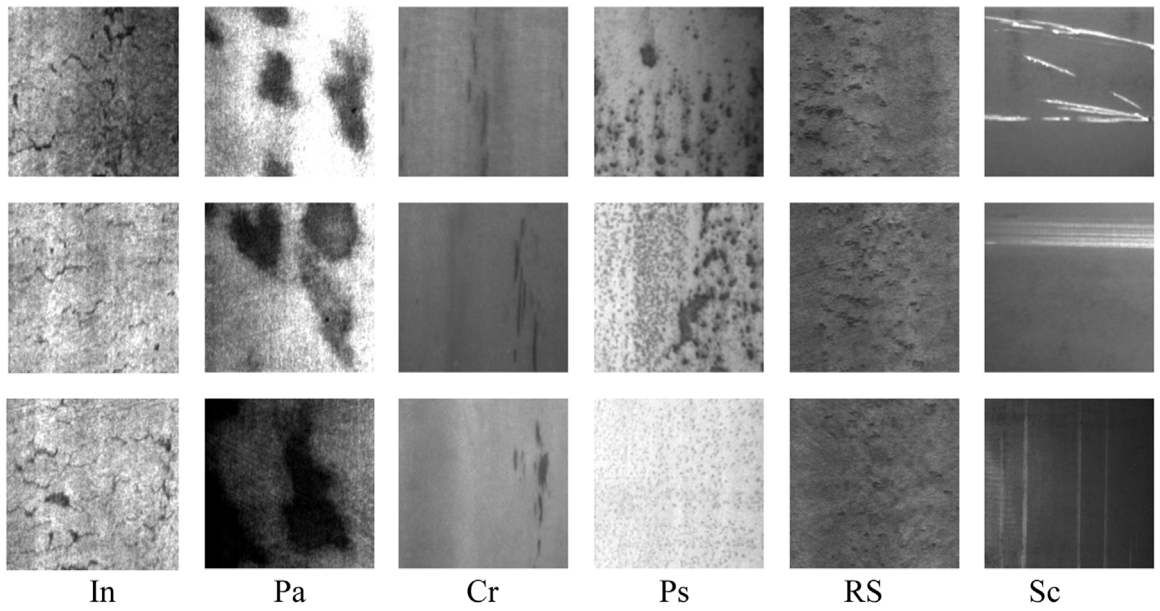


FIGURE 5
Six types of steel surface defects.

TABLE 1 Average classification accuracy of different sample sizes.

Sample size for each category (total sample size)	Average accuracy (%)		Increase (%)
	ACGAN	Ours	
200 (1,200)	94.83	**98.67**	4.04
150 (900)	93.67	**98.33**	4.66
100 (600)	90.00	**95.50**	5.5
50 (300)	83.83	**94.67**	10.84
30 (180)	77.83	**94.00**	16.17
10 (60)	66.50	**89.67**	23.17

Bold values mean the best results.

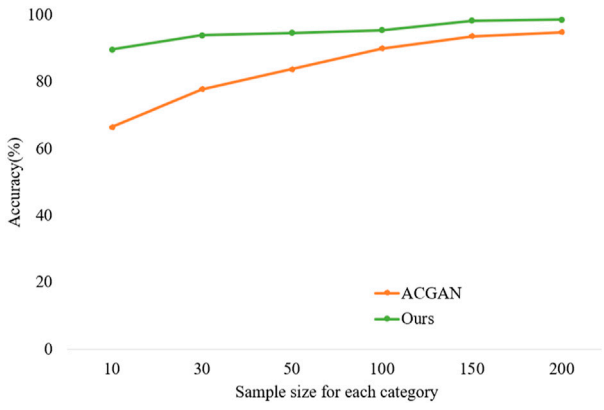


FIGURE 6
Trend chart of classification results under different sample sizes.

effective than the model in this paper when dealing with few-shot problems.

To illustrate the classification ability of the proposed model for each type of defect, Figure 7 shows the confusion matrix of our model under different sample sizes, where the numbers 0–5 in the abscissa and ordinate represent defect types, respectively: Cr, In, Pa, PS, RS, and Sc. It is evident that our method can train an ideal model under different training sample sizes and can accurately classify most of the defects. Moreover, when the sample size is 200, the model can accurately classify all Pa defects. Under different training sample sizes, the cases of classifying Cr as RS and RS as Cr occupy a large proportion in the wrong classification cases. The high similarity between Cr and RS defects and the lack of distinct inter-class features lead to misjudgment of the model. The overall results demonstrate that the method proposed in this paper only misjudges a few fault types under different sample sizes, and the overall accuracy remains high as the sample size decreases.

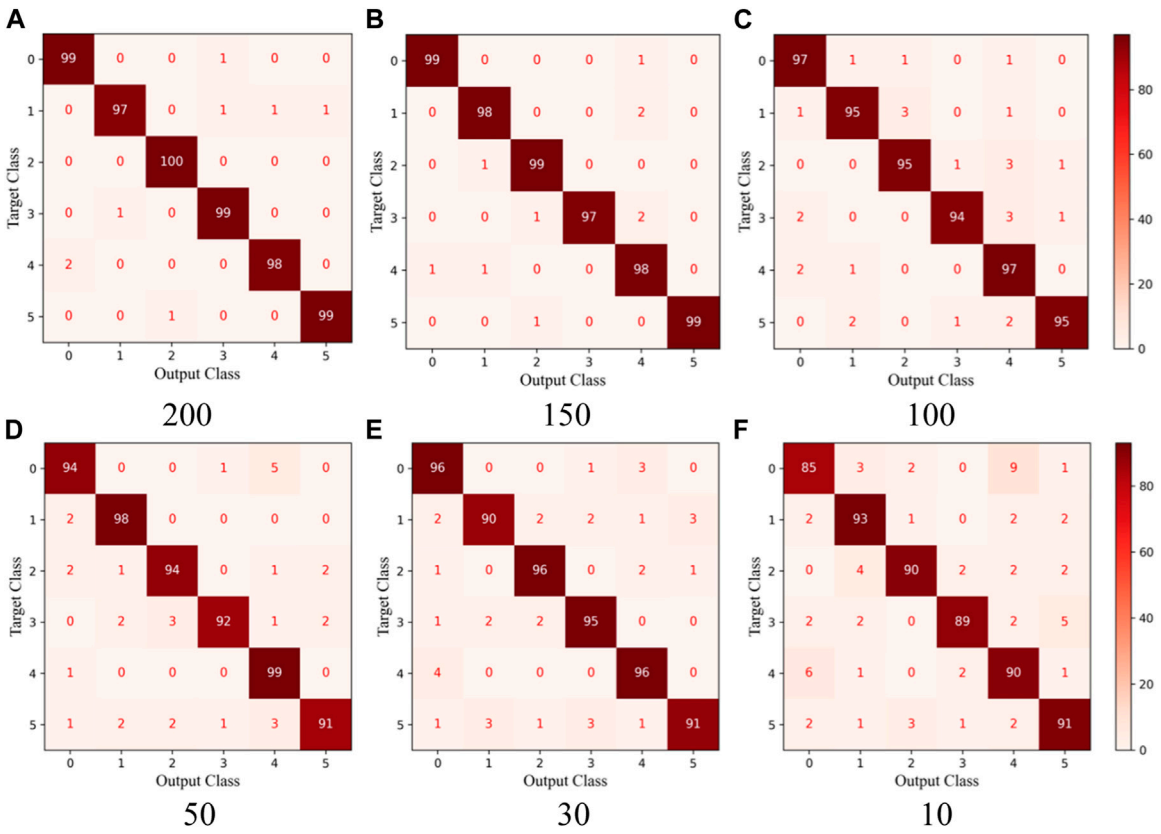


FIGURE 7
Confusion matrix of our model under different training sample sizes. **(A)** 200, **(B)** 150, **(C)** 100, **(D)** 50, **(E)** 30, and **(F)** 10.

TABLE 2 Results of steel surface defects under different methods and sample sizes.

Methods	Average accuracy (%)					
	200	150	100	50	30	10
ResNet18	92.33	90.67	85.00	83.33	76.33	59.5
ResNet50	93.00	92.33	85.33	83.00	77.00	63.17
Res-ACGAN	96.17	95.00	91.00	84.67	79.00	70.50
[28]	96.50	95.50	91.00	89.50	87.50	76.33
[35]	96.67	94.67	90.50	85.33	84.83	71.33
Ours (lack MSFE)	97.00	96.00	94.33	93.67	91.33	86.00
Ours	**98.67**	**98.33**	**95.50**	**94.67**	**94.00**	**89.67**

Bold values mean the best results.

In order to further validate the classification performance of our model, we compare it with the classic ResNet18 and ResNet50 classification methods. To ensure the efficient classification performance of the classic classification models, the ResNet18 and ResNet50 models are pre-trained using the ImageNet dataset. Additionally, we also compared with the latest few-shot deep learning classification models, including: the model proposed by Lian et al. [28], which combine generative adversarial networks and convolutional neural networks to generate exaggerated defect image samples to ensure the accuracy of micro-surface defect detection; and the model proposed by Li et al. [35], which replace the fully connected classification layer with an orthogonal SoftMax layer, significantly reducing the complexity of the model and making it suitable for few-shot classification. Moreover, in order to fully demonstrate the impact of MSFE on the final classification results, MSFE is deliberately excluded in the original framework and a corresponding experiment is conducted. The experimental results are presented in Table 2, and it can be seen that, in the case of different sample sizes, the methods proposed in this paper have achieved the best results and achieved the highest classification accuracy.

In order to further verify the importance of introducing various parts in our model, we compare the classification performance of the model after introducing the residual module, Wasserstein distance and penalty weight GP [30], Wasserstein divergence, and MSFE into the ACGAN model respectively. At the same time, in order to verify the important role played by the residual module in the discriminator network, we replace the residual module in the discriminator of our model with the CBAM attention mechanism module proposed by [35], and introduce the SENet module to conduct comparative experiments. The training sample size is 200, and the experimental results are shown in Table 3. It can be found that after adding the residual module to the original model, the classification accuracy of ACGAN increases by 1.34%, the

TABLE 3 Classification accuracy of introducing different modules.

Method	Accuracy (%)
ACGAN	94.83
ACGAN + Res	96.17
ACGAN + Wasserstein + GP	95.33
ACGAN + Wasserstein-div	95.83
ACGAN + Res + Wasserstein + GP	96.50
ACGAN + SENet + Wasserstein + GP	95.83
ACGAN + Res + SENet + Wasserstein + GP	96.67
ACGAN + Res + CBAM + Wasserstein + GP	96.83
ACGAN + Res + Wasserstein-div	97.00
ACGAN + Res + Wasserstein-div + MSFM (Ours)	**98.67**

Bold values mean the best results.

classification accuracy of ACGAN + Wasserstein + GP increases by 1.17%, and the classification accuracy of ACGAN + SENet + Wasserstein + GP increases by 0.84%. After adding Wasserstein divergence to the original model, the classification accuracy of ACGAN increases by 1%, and the classification accuracy of ACGAN + Res increases by 0.83%, which is higher than that of using Wasserstein distance and penalty weight, showing that introducing the residual module and Wasserstein divergence into the model can improve the feature extraction ability of the model and further improve the model's ability to discriminate and classify sample images. In addition, the introduction of attention mechanism modules SENet and CBAM in the discriminator network can improve the classification ability of the model, but the discriminator network structure proposed in this paper has achieved the best results in experiments. The incorporation of MSFE in the original framework results in a 1.67% increase in classification accuracy. This implies that employing semantic features at varying levels to guide the generator can enhance its efficiency, thereby advancing the classification abilities of the discriminator.

4.3 Quality assessment of generated samples

Figure 8 presents a comparison of steel surface defect samples generated by different models, including ACGAN, the model augmented with SENet module, the model augmented with CBAM module [35], the proposed method while lacking MSFE, and our model. The training process utilizes 200 samples of each type of defect, with 10,000 iterations and other parameters hold constant.

It can be observed that, compared to the original sample in Figure 5, the samples generated by the method proposed in this

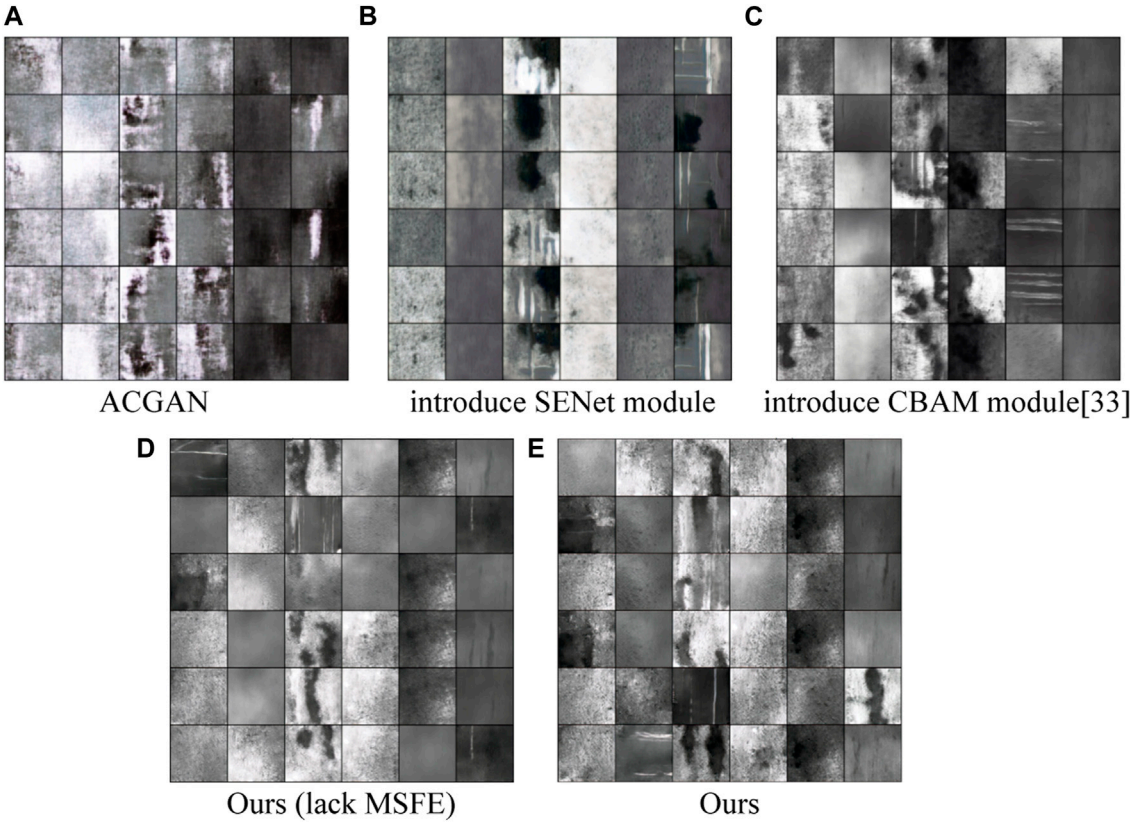


FIGURE 8
Sample images generated by different methods. **(A)** ACGAN. **(B)** introduce SENet module. **(C)** introduce CBAM module [34]. **(D)** Ours (lack MSFE). **(E)** Ours.

TABLE 4 Comparison of MSE and SSIM values of different models.

Methods	MSE	SSIM
ACGAN	347.8543	0.6935
SENet-ACGAN	272.5821	0.7074
CBAM-ACGAN	222.4989	0.7583
Ours (lack MSFE)	193.8484	0.7828
Ours	**184.2617**	**0.7912**

Bold values mean the best results.

paper are more distinct and the quality of the samples are also much better. For instance, for the defect of scratch, such as the third one in the fifth row and the second one in the sixth row in Figure 8E generated by the method proposed in this paper, when compared to the last one in the second row in (a), the last one in the first row in (b), the third one in the fourth row in (c), and the last one in the last row in (d), its defect features are more discernible, the defect is sharper, and it is also more similar to the original sample image. Although the version of lacking MSFE can also generate high-quality sample images, it is evident that its feature extraction ability is inadequate, leading to blurred images and unclear semantic information, as demonstrated in Figure 8D, specifically in the fifth one of the second row and the second item of the fourth row.

In order to assess the quality of samples generated by different models, the MSE (Mean Square Error) and SSIM (Structural Similarity) metrics are employed to evaluate the sample quality. MSE is a metric that reflects the degree of discrepancy between the estimator and the estimated quantity; SSIM is used to measure the similarity between two images. The results of different models are presented in Table 4. The smaller the value of MSE, or the larger the value of SSIM, the larger the similarity between original image and generated image. It can be seen from Table 4 that the MSE and SSIM of our model are more proximate to the original images than other methods, which demonstrates that the sample data distribution generated by our model is more similar to the original sample distribution, and also shows that MSFE and Wasserstein divergence can improve the quality of samples generated by the model.

5 Conclusion

Aiming at the difficulties of steel surface defect few-shot classification, this paper introduces multi-level semantic feature extractor under the residual adversarial learning network framework to generate high-quality samples and achieves promising steel surface defect classification. First, we modify the network structure of the adversarial learning model by the residual module, so that the model can obtain more information during training and generate synthetic data to the original sample. To overcome the challenge of inadequate feature extraction in generator networks which may lead to suboptimal sample quality in small-sample environments, we design a multi-level semantic feature extractor for obtaining diverse semantic information at various levels. By leveraging this comprehensive semantic information, we directed sample generation. At the same time, the Wasserstein divergence is introduced into the loss function to solve the problem of unstable model training and to improve the generation efficiency and classification performance of the model. Experiments are conducted on the steel surface defect dataset NEU-CLS from Northeastern University. The results demonstrate that, under the condition of the restricted number of training samples, the method proposed in this paper achieves the highest classification accuracy. Moreover, when the number of training data is reduced, our method exhibits better stability and robustness than classical classification models and state-of-the-art of deep learning models. Additionally, in terms of the quality of generated samples, the MSE value and SSIM value of the samples generated by the model proposed in this paper are the closest to the original samples, further showing the effectiveness of our proposed method. With the popularization of sensors and lightweight devices, the demand for model compression and lightweight models is becoming increasingly important. Improving the real-time performance of defect detection systems is the main trend for deploying online detection systems in actual industrial production in the future.

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

LH: Ideas, methodology, experimental design, formal analysis. PS: Software, validation, data curation. ZP: Supervision, Writing—review and editing. YX: Supervision, writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Establishment of Key Laboratory of Shenzhen Science and Technology Innovation Committee under Grant ZDSYS20190902093015527, the Shenzhen Science and Technology Innovation Committee under Grant JSGG20220831104402004, and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

Conflict of interest

Authors PS and ZP were employed by HBIS Digital 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

References

1. Ayarkwa J. Influence of wood defects on some mechanical properties of two tropical Ghanaian hardwoods. *J Ghana Sci Assoc* (1999) 1:131–47. doi:10.4314/jgsa.v1i2.17813

2. Yu Z, Wu X, Gu X. *Fully convolutional networks for surface defect inspection in industrial environment*. Cham: Springer (2017).

3. Song G, Song K, Yan Y. EDRNet: Encoder–Decoder residual network for salient object detection of strip steel surface defects. *IEEE Trans Instrumentation Meas*, 2020, 69:1–. doi:10.1109/TIM.2020.3002277

4. Zaghdoudi R, Seridi H, Ziani S. Binary Gabor pattern (BGP) descriptor and principal component analysis (PCA) for steel surface defects classification[C]. In: Proceeding of the 2020 International Conference on Advanced Aspects of Software Engineering (ICAASE); November 2020. IEEE (2020). p. 1–7.

5. Hu H, Li Y, Liu M, Liang W. Classification of defects in steel strip surface based on multiclass support vector machine. *Multimedia tools Appl* (2014) 69(1):199–216. doi:10.1007/s11042-012-1248-0

6. Duan C, Zhang T. Two-stream convolutional neural network based on gradient image for aluminum profile surface defects classification and recognition. *IEEE Access* (2020) 8:172152–65. doi:10.1109/access.2020.3025165

7. Liu X, He W, Zhang Y, Yao S, Cui Z. Effect of dual-convolutional neural network model fusion for Aluminum profile surface defects classification and recognition. *Math Biosciences Eng* (2022) 19(1):997–1025. doi:10.3934/mbe.2022046

8. Mayr M, Hoffmann M, Maier A, Christlein V. Weakly supervised segmentation of cracks on solar cells using normalized L p norm[C]. In: Proceeding of the 2019 IEEE International Conference on Image Processing (ICIP); September 2019. IEEE (2019). p. 1885–9.

9. Huang Y, Qiu C, Yuan K. Surface defect saliency of magnetic tile. *Vis Comp* (2020) 36(1):85–96. doi:10.1007/s00371-018-1588-5

10. Li K, Qi Y, Su L, Gu J, Su W. Visual inspection of steel surface defects based on improved auxiliary classification generation adversarial network[J/OL]. *Chin J Mech Eng* (2023) 1–9. Available at: http://kns.cnki.net/kcms/detail/11.2187.TH.20220526.1827.106.html.

11. Panaretos VM, Zemel Y. Statistical aspects of Wasserstein distances. *Annu Rev Stat its Appl* (2019) 6:405–31. doi:10.1146/annurev-statistics-030718-104938

12. Radford A, Metz L, Chintala S. *Unsupervised representation learning with deep convolutional generative adversarial networks[J]*. arXiv preprint (2015).

13. Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans[C]. In: *International conference on machine learning*. New York: PMLR (2017). p. 2642–51.

14. Dosovitskiy A, Brox T. Generating images with perceptual similarity metrics based on deep networks. *Adv Neural Inf Process Syst* (2016) 29.

15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* (2020) 63(11):139–44. doi:10.1145/3422622

16. Lu HP, Su CT. CNNs combined with a conditional GAN for mura defect classification in TFT-LCDs. *IEEE Trans Semiconductor Manufacturing* (2021) 34(1):25–33. doi:10.1109/tsm.2020.3048631

17. Li X, Zhang W, Ding Q. Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks. *IEEE Trans Ind Elect* (2018) 66(7):5525–34. doi:10.1109/TIE.2018.2868023

18. Liu J, Zhang BG, Li L. Defect detection of fabrics with generative adversarial network based flaws modeling[C]. In: Proceeding of the 2020 Chinese Automation Congress (CAC); November 2020; Shanghai, China. IEEE (2020). p. 3334–8.

19. Cheon S, Lee H, Kim CO, Lee S. Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. *IEEE Trans Semiconductor Manufacturing* (2019) 32(2):163–70. doi:10.1109/tsm.2019.2902657

20. Nakazawa T, Kulkarni DV. Wafer map defect pattern classification and image retrieval using convolutional neural network. *IEEE Trans Semiconductor Manufacturing* (2018) 31(2):309–14. doi:10.1109/tsm.2018.2795466

21. Zhu H, Ge W, Liu Z. Deep learning-based classification of weld surface defects. *Appl Sci* (2019) 9(16):3312. doi:10.3390/app9163312

22. Wan X, Zhang X, Liu L. An improved VGG19 transfer learning strip steel surface defect recognition deep neural network based on few samples and imbalanced datasets. *Appl Sci* (2021) 11(6):2606. doi:10.3390/app11062606

23. Han T, Liu C, Yang W, Jiang D. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. *ISA Trans* (2020) 97:269–81. doi:10.1016/j.isatra.2019.08.012

24. Liu Y, Yuan Y, Liu J. Deep learning model for imbalanced multi-label surface defect classification. *Meas Sci Tech* (2021) 33(3):035601. doi:10.1088/1361-6501/ac41a6

25. Jain S, Seth G, Paruthi A, Yang EWR, Lwin S, Yeo TT, et al. Pseudoaneurysm resulting in rebleeding after evacuation of spontaneous intracerebral hemorrhage. *J Intell Manufacturing* (2020) 143:1–6. doi:10.1016/j.wneu.2020.07.088

26. He Y, Song K, Dong H, Yan Y. Semi-supervised defect classification of steel surface based on multi-training and generative adversarial network. *Opt Lasers Eng* (2019) 122:294–302. doi:10.1016/j.optlaseng.2019.06.020

27. Zhao Z, Li B, Dong R, Zhao P. A surface defect detection method based on positive samples[C]. In: *Pacific rim international conference on artificial intelligence*. Cham: Springer (2018). p. 473–81.

28. Lian J, Jia W, Zareapoor M, Zheng Y, Luo R, Kumar D. Deep-learning-based small surface defect detection via an exaggerated local variation-based generative adversarial network. *IEEE Trans Ind Inform* (2019) 16(2):1343–51. doi:10.1109/TII.2019.2945403

29. Tsai DM, Fan SKS, Chou YH. Auto-annotated deep segmentation for surface defect detection. *IEEE Trans Instrumentation Meas* (2021) 70:1–10. doi:10.1109/tim.2021.3087826

30. Dumoulin V, Visin F. *A guide to convolution arithmetic for deep learning* (2016). arXiv preprint arXiv:1603.07285.

31. Arjovsky M, Chintala S, Bottou L. *Wasserstein GAN* (2017).

32. Wu J, Huang Z, Thoma J, Acharya D, Gool L. Wasserstein divergence for gans[C]. In: *Proceedings of the European conference on computer vision*. Switzerland: ECCV (2018). p. 653–68.

33. Dong H, Song K, He Y, Xu J, Yan Y, Meng Q. PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection. *IEEE Trans Ind Inform* (2019) 16(12):7448–58. doi:10.1109/TII.2019.2958826

34. Meng Z, Li Q, Sun D, Cao W, Fan F. An intelligent fault diagnosis method of small sample bearing based on improved auxiliary classification generative adversarial network. *IEEE Sensors J* (2022) 22(20):19543–55. doi:10.1109/jsen.2022.3200691

35. Li X, Chang D, Ma Z, Tan ZH, Xue JH, Cao J, et al. OSLNet: Deep small-sample classification with an orthogonal softmax layer. *IEEE Trans Image Process* (2020) 29:6482–95. doi:10.1109/tip.2020.2990277

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 04 July 2023
DOI 10.3389/fphy.2023.1214206

OPEN ACCESS

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

REVIEWED BY
Guanqiu Qi,
Buffalo State College, United States
Sheng Xiang,
Chongqing University, China

*CORRESPONDENCE
Jin Duan,
duanjin@vip.sina.com

RECEIVED 29 April 2023
ACCEPTED 14 June 2023
PUBLISHED 04 July 2023

CITATION
Duan J, Zhang H, Liu J, Gao M, Cheng C and Chen G (2023), A dual-weighted polarization image fusion method based on quality assessment and attention mechanisms.
Front. Phys. 11:1214206.
doi: 10.3389/fphy.2023.1214206

A dual-weighted polarization image fusion method based on quality assessment and attention mechanisms

Jin Duan*, Hao Zhang, Ju Liu, Meiling Gao, Cai Cheng and Guangqiu Chen

College of Electronic Information Engineering, Changchun University of Science and Technology, Changchun, China

This paper proposes a dual-weighted polarization image fusion method based on quality assessment and attention mechanisms to fuse the intensity image (S0) and the degree of linear polarization (DoLP). S0 has high contrast and clear details, and DoLP has an outstanding ability to characterize polarization properties, so the fusion can achieve an effective complementation of superior information. We decompose S0 and DoLP into base layers and detail layers. In the base layers, we build a quality assessment unit combining information entropy, no-reference image quality assessment, and local energy to ensure the fused image has high contrast and clear and natural visual perception; in the detail layer, we first extract depth features using the pre-trained VGG19, then construct an attention enhancement unit combining space and channels, and finally effectively improve the preservation of detail information and edge contours in the fused image. The proposed method is able to perceive and retain polarization image features sufficiently to obtain desirable fusion results. Comparing nine typical fusion methods on two publicly available and own polarization datasets, experimental results show that the proposed method outperforms other comparative algorithms in both qualitative comparison and quantitative analysis.

KEYWORDS

image fusion, polarization image, double weighting, quality assessment, attention mechanisms

1 Introduction

Image fusion techniques aim to synthesize images by fusing complementary information from multiple source images captured by different sensors [1]. In recent years, many fusion methods have been proposed. According to [2], the classical fusion methods mainly include multi-scale transform-based methods [3, 4], sparse representation-based methods [5, 6], and neural network-based methods [7, 8]. Most of these methods mainly involve three key operations, such as image transformation, activity level measurement, and fusion strategy design. However, since these operations are mainly designed in a manual way, they may not be suitable for different situations, thus limiting the accuracy of the fusion results.

Fusion methods for deep learning [9, 10] have been widely studied and applied, with better fusion effects than traditional methods by virtue of their powerful feature extraction capabilities. Some scholars have shown that the combination of CNN with various traditional fusion methods not only has outstanding effects but can also effectively

reduce the workload and save computational resources by taking advantage of the migratory nature of CNN-encoded information. For example, Li et al. [11] proposed a fusion algorithm using ResNet50 and ZCA with a weighted average strategy to reconstruct the fused images, which significantly improved the fusion effect of the images. However, since they designed a simple fusion strategy, it may lead to the problem of insufficient information combination during the fusion process. Li et al. [12] used a densely connected network combined with an attention mechanism at the same time to enable the network to better capture the structural information of the source image. However, this method does not take into account the information differences at different scales, so the fused image may suffer from the loss of detailed information. Meanwhile, both methods are for the fusion task of infrared images, and the fusion effect is not satisfactory when applied directly to polarized images. At present, there are relatively few studies on polarization image fusion. Wang et al. [13] combined NSCT and CNN to propose a polarization image fusion network. Although it has some enhancement effect, conventional strategies such as weighted average and local energy are still used in the fusion process, and there is no analysis of polarization images or a more reasonable strategy design.

To solve the problems of the above methods, we propose a dual-weighted polarization image fusion method based on quality assessment and attention mechanisms. The S0 and DoLP images are decomposed into base layers and detail layers, and different strategies are constructed for fusion, respectively. Among them, S0 can reflect the spectral information of the object and is mainly used to describe the reflectance and transmittance; DoLP can reflect the difference in polarization characteristics between different material substances and provide information such as surface shape, shadow, and roughness. The fusion of S0 and DoLP can make up for the disadvantage that S0 cannot provide sufficient information in certain scenes and thus improve the target recognition ability in complex backgrounds. The implementation process of our method mainly includes: in the base layer, a quality assessment unit is constructed to achieve a balanced and reasonable fusion effect. Through comprehensive assessment and calculation of image quality, the best fusion relationship between the base layers can be obtained, and then a clear and natural fusion base layer can be obtained; in the detail layers, the depth features are first extracted using the pre-trained VGG19, and then an attention enhancement unit is constructed to enhance the polarization image detail layers from different dimensions, which can effectively combine global contextual information and improve the structural features of the fused detail layer. Using the fused base layer and the detail layer for reconstruction, the obtained fused images have rich texture details with high enough contrast and natural visual perception. The main contributions to this paper are as follows.

1. Dual-weighted fusion method. We propose a dual-weighted polarization image fusion method with strong perception and retention of features of polarization images, which is more suitable for the fusion task of polarization images.
2. Quality assessment-based fusion strategy. We construct a quality assessment unit consisting of information entropy, no-reference image quality assessment, and local energy. The optimal fusion weight is obtained by assessing the information quality of the base layers, which is used to ensure the high contrast and natural visual perception of the fused image.
3. Attention enhancement-based fusion strategy. We construct an attention enhancement unit consisting of space and channels to enhance the global features of the detail layers in two dimensions, which can effectively improve the detail information and texture contours to obtain a fused image that fully combines intensity information and polarization characteristics.

The rest of this paper is organized as follows. Section 2 briefly reviews the research related to image fusion methods based on multiscale transforms and attention mechanisms. In Section 3, the details of our proposed method are described. Section 4 conducts experiments on the public dataset and our polarization dataset, and the experimental results are analyzed. The paper is summarized in Section 5.

2 Related work

2.1 Multiscale transform for image fusion

The fusion method based on multiscale transformation has the advantages of simplicity, efficiency, and outstanding effect compared with other methods, so it is most widely studied and applied. The main implementation process is to first decompose the source image into several different scales, then fuse the images of different scales according to specific fusion rules, then perform the corresponding multi-scale inverse transform, and finally obtain the fused image. Usually, many methods divide the image into high-frequency and low-frequency parts, or basis and detail parts. Among them, the low-frequency part and the basis part represent the energy distribution of the image, and the commonly used fusion rules include average value, local energy maximum, etc.; the high-frequency part and the detail part represent the edge and detailing features of the image, and the fusion rules include absolute value maximum, adaptive weighting, etc. Since the fusion strategy has a great influence on the fusion effect, it is important and one of the most challenging studies to design a more reasonable strategy to improve the fusion effect.

Fusion methods based on multiscale transformations have been widely studied in recent years. Wang et al. [13] proposed a fusion algorithm for polarized images. Noise removal and pre-fusion processing are performed first, and then the polarization and intensity images obtained by pre-fusion are decomposed by NSCT, and then the fusion strategies for high and low frequencies are developed separately, and finally the target fusion image is obtained by inverse transformation. Zhu et al. [14] proposed a fusion method based on image cartoon-texture decomposition and sparse representation. After decomposing the source image into cartoon and texture parts, the proposed spatial morphological structure preservation method and the sparse representation-based method are used for fusion, respectively. Zhu et al. [15] proposed a multimodal medical image fusion method. The high-pass and low-pass subbands were fused using phase congruency-based and local Laplace energy-based rules,

respectively, and the effectiveness of the proposed method was verified experimentally. Li et al. [16] proposed a fusion framework that decomposes the source image into a base part and a detail part. Among them, the base part uses a weighted average fusion strategy, and the detail part uses a maximum selection strategy to fuse the extracted multilayer depth features. Finally, the fused base and detail parts are combined to obtain a clear and natural fused image; Liu et al. [17] proposed an infrared polarization image fusion method. A multi-decomposition latent low-rank representation is used to decompose the source image into low-rank and significant parts, and different strategies are used to process the weight map, and finally the fused image is reconstructed. Hu et al. [18] proposed an improved hybrid multiscale fusion algorithm. The image is first decomposed into low-frequency and high-frequency parts using the support value transform, and then the prominent edges are further extracted from these support value images using the shearlet transform of NSST. Zou et al. [19] proposed a visible and near-infrared image fusion method based on a multiscale gradient guided edge-smoothing model and local gradient weighting, which has obvious advantages in maintaining edge details and color naturalness.

It can be found that the fusion rule of existing methods rarely analyzes the images, and most of them still use manually designed rules that do not consider the differences between images. To solve these problems, we assess the image quality and combine the attention mechanism to design two novel fusion strategies and apply them to different layers of polarization images. Among them, the quality assessment unit is applied to the base layers, which can obtain the best fusion weight based on the image quality, and the attention enhancement unit is applied to the detail layers, which can enhance the detail features by combining global contextual information.

2.2 Attention mechanism for image fusion

The attention mechanism is consistent with the human visual system and has been widely studied because it can better perceive and extract image features [20, 21]. The purpose of fusion is to combine the superior information from different images, and more weight needs to be given to salient parts during the fusion process, such as features like detailed textures and edge contours. The ability to maintain the integrity of the salient target regions using attention mechanisms can effectively improve the quality of fused images.

In recent years, many attention-based or saliency-based fusion methods have been proposed. Wang et al. [22] proposed a two-branch network based on an attention mechanism in the fusion block while using an attention model with large perceptual fields in the decoder to effectively improve the quality of fused images. Li et al. [23] proposed a generative adversarial based on multiscale feature migration and a deep attention mechanism for the fusion of infrared and visible images and achieved excellent fusion results. Liu et al. [24] designed a two-stage enhancement framework based on attention mechanisms and a feature-linking model with the advantage of being able to suppress noise effectively. Zhang et al. [25] proposed an iterative visual saliency map to retain more details of the infrared image and calculate the weight map based on the designed multiscale bootstrap filter and saliency map, which in turn guides the texture fusion. Cao et al. [26] proposed a fusion network based on multi-scale and attention mechanisms, and the advantages of the proposed method were verified by experiments on two datasets. Wang et al. [27] proposed a multimodal image fusion framework that was designed mainly using multiscale gradient residual blocks and a pyramid split attention module.

At present, there is no polarization image fusion method that extracts features using pre-trained CNNs in a multi-scale layer while combining an attention mechanism. Specifically, we combine a deep learning framework with an attention mechanism. A pre-trained VGG19 is used to extract the depth features of the detail layers, while an attention enhancement unit consisting of space and channels is constructed to enhance the global features of the detail layers in two dimensions, which is used to improve the detail information and texture contours of the fused images.

3 Proposed method

Our method consists of three parts, and the framework is shown in Figure 1.

(1) Decomposition: the method of literature [16] is used to decompose S0 and DoLP into base layers and detail layers.
(2) Fusion: for the base layers, the fused base layer is obtained by weighting calculation using the quality assessment unit; for the detail layers, the fused detail layer is obtained by using the attention enhancement unit.
(3) Reconstruction: the fusion polarization images are reconstructed using the fused base layers and the detail layers, which can be formulated as in Eq. 1.

$$F(x, y) = F_b(x, y) + F_d(x, y) \tag{1}$$

where $F_b(x, y)$ and $F_d(x, y)$ denote the fused base layer and the detail layer, respectively; and $F(x, y)$ denotes the fused image.

3.1 QAU for base layer fusion

In the fusion process of the base layers, it is necessary to consider how to reasonably retain the advantageous information of different source images. Therefore, we designed a quality assessment unit (QAU) consisting of information entropy (EN) [28], no-reference image quality assessment (NR-IQA) [29], and local energy (LE) [30]. The QAU is shown in Figure 2.

Among them, EN [31] can both reflect the amount of information in the image and serve as an evaluation index of the fused image, as shown in Eq. 2. In general, the more information the image contains, the larger the EN value. Since DoLP has more noise compared to S0, the EN value of DoLP will be higher. Therefore, it is not accurate enough if only EN is used to evaluate the image quality. Image quality assessment (IQA) can evaluate the quality of the information contained in the source image, but since high-quality original images are more difficult to obtain, we use the no-reference model (NR) instead of the full-reference model. Image quality is judged by the NR-IQA

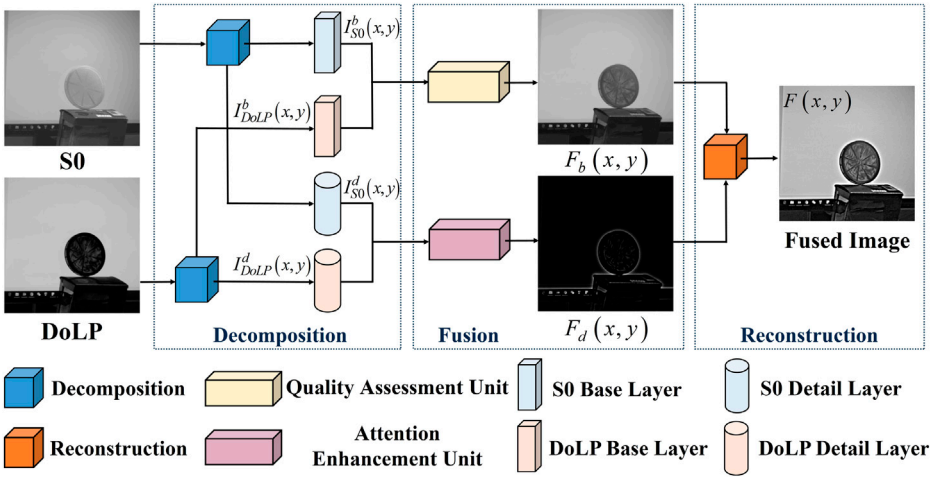


FIGURE 1
The framework of the proposed method. $I^b_{S0}(x,y)$ and $I^b_{DoLP}(x,y)$ denote the S0 and DoLP base layers, respectively, while $I^d_{S0}(x,y)$ and $I^d_{DoLP}(x,y)$ denote the detail layers. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

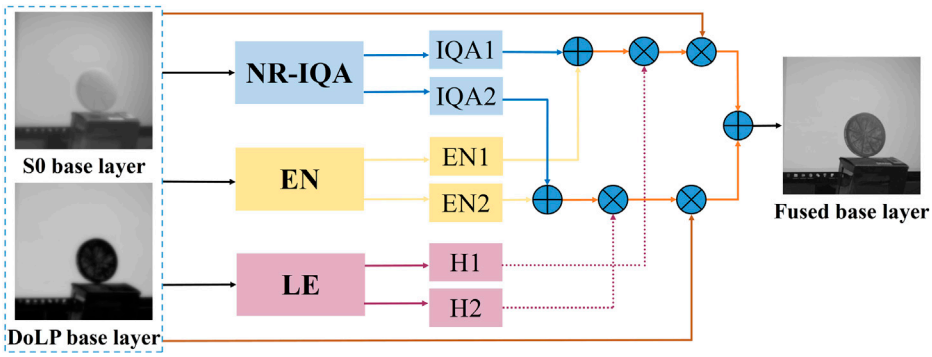


FIGURE 2
The procedure of the fusion strategy for base layers. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

simulation of the visual system, which measures the degree of distortion caused by block effects, noise, compression, etc., in each semantic region of the base layers. Since S0 consists of specular and diffuse reflections, which represent the sum of the intensities of two orthogonal polarization directions, NR-IQA will calculate a higher value. Also, because DoLP has a larger microscopic surface difference, NR-IQA will judge it as a low-quality image. If only NR-IQA is used to assess image quality, the balance of information relationships cannot be accurately calculated. Therefore, we combine NR-IQA with EN to be able to estimate the weight relationship of the source image substrate more reasonably and retain more information. In addition, since LE can reflect the degree of uniformity of image gray distribution, it is used as an adjustment factor to ensure that the fused base layer has a natural visual perception. The formulas are described as Eq. 3, Eq. 4, respectively.

$$EN = -\sum_{l=0}^{L-1} P_l \log_2 P_l \tag{2}$$

$$E_k(x,y) = \sum_{m=-(M-1)/2}^{(M-1)/2} \sum_{n=-(N-1)/2}^{(N-1)/2} w(m,n) \times \left[D_k(x+m, y+n)\right]^2 \tag{3}$$

$$H_k(x,y) = \frac{E_k(x,y)}{\sum_{k=1}^{n} E_k(x,y)} \tag{4}$$

where L is the number of gray levels, which is set to 256, and P_l denotes the probability of each level. $n = 2$, denotes S0 and DoLP, respectively; $M \times N$ and $w(m,n)$ are the window area and coefficients, respectively; (x, y) denotes the pixel centroid; and $D_k(x, y)$ denotes the coefficient value of the source image at that point.

Therefore, by combining the above three quality assessment methods, the optimal weight map can be obtained, which is defined as $M_k(x, y)$. Where $EN(\cdot)$ represents the calculation of information entropy, $NR\text{-}IQA(\cdot)$ is the image quality assessment without reference, and $H_k(\cdot)$ is the adjustment factor obtained from LE. The formula is as follows.

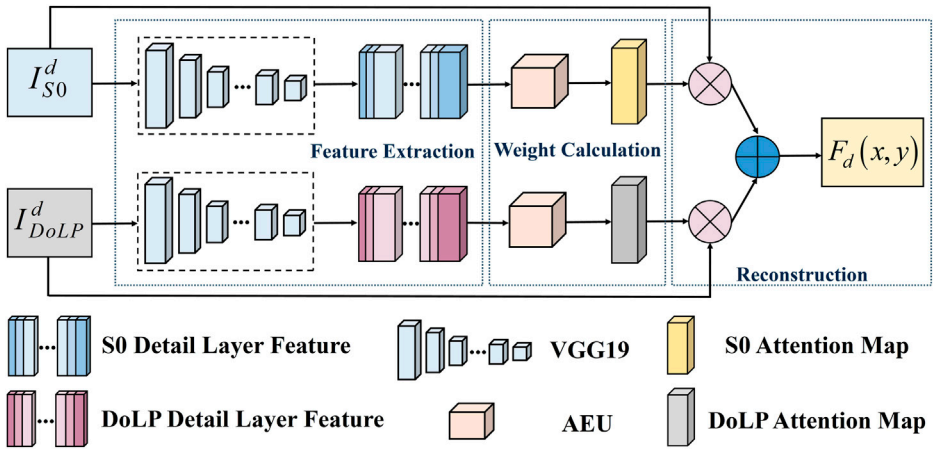


FIGURE 3
The procedure of the fusion strategy for detail layers.

$$M_k(x,y) = H_k\left(I_k^b(x,y)\right) \cdot \left(EN\left(I_k^b(x,y)\right) + NR\text{-}IQA\left(I_k^b(x,y)\right)\right) \quad (5)$$

The base layers of S0 and DoLP are input to QAU to obtain the corresponding weight maps, and the weighting is calculated to obtain the fused base layer with the following equation.

$$F_b(x,y) = \sum_{k=1}^{n} M_k(x,y) \cdot I_k^b(x,y) \quad (6)$$

3.2 AEU for detail layer fusion

As shown in Figure 3, the detail layer fusion process consists of three main parts: feature extraction, weight calculation, and reconstruction. First, we extract the detail layer depth features of S0 and DoLP using the pre-trained VGG19. Second, the weight map is obtained using the attention enhancement unit. And then, the weight map is scaled to the source image size. Among them, the reasons for using the pre-trained VGG19 are analysed in Section 4.3.2. The weight map is obtained from the activation level map by calculating the soft-max operator, defined as Eq. 7.

$$W_k^i(x,y) = \frac{C_k^i(x,y)}{\sum_{k=1}^{n} C_k^i(x,y)} \quad (7)$$

Finally, the weighting calculation is performed to obtain the fused detail layer, which can be formulated as in Eq. 8.

$$F_d(x,y) = \sum_{k=1}^{n} W_k^i(x,y) \cdot I_k^d(x,y) \quad (8)$$

where $i \in \{1,2,3,4\}$, represent *relu*_1_1, *relu*_2_1, *relu*_3_1 and *relu*_4_1, respectively.

As shown in Figure 4, the attention enhancement unit (AEU) consists of spatial attention (SA) and channel attention (CA), which aim to enhance the semantic targets and texture contours in the detail layers of the source image. The AEU can extract the feature distributions in the source image that complement each other and can generate different weights for spatial and channel features, while the global information of the source image can be enhanced, which in turn improves the feature representation of the fused image. Among them, the SA focuses more on information such as high-frequency regions, which can enhance the details of the fused image, while the CA focuses on different channel features with completely different weighting information.

The SA consists of L1-norm and soft-max, and the formula is formulated as follows.

$$\alpha_k^i(x,y) = \frac{\left\|F_k^i(x,y)\right\|_1}{\sum_{k=1}^{n}\left\|F_k^i(x,y)\right\|} \quad (9)$$

$$\phi_k^i(x,y) = \alpha_k^i(x,y) \times F_k^i(x,y) \quad (10)$$

where $F_k^i(x,y)$ denotes the feature vector and $\|\cdot\|_1$ denotes the L1 parametric calculation.

The CA consists of a global average pooling operator ($P(\cdot)$) and soft-max, and the formula is as follows.

$$\eta_k^i(x,y) = P(F_k^i(x,y)) \quad (11)$$

$$\beta_k^i(x,y) = \frac{\eta_k^i(x,y)}{\sum_{k=1}^{n}\eta_k^i(x,y)} \quad (12)$$

$$\psi_k^i(x,y) = \beta_k^i(x,y) \times F_k^i(x,y) \quad (13)$$

The detail layer depth features of S0 and DoLP are fed into SA and CA, and the corresponding weight maps are obtained and then summed, as defined in the following equation.

$$C_k^i(x,y) = SA(F_k^i(x,y)) + CA(F_k^i(x,y)) \quad (14)$$

4 Results and analysis

4.1 Experiment settings

We compare nine algorithms on two publicly available [32, 33] datasets and our own. Among them, the public datasets are from the University of Tokyo and King Abdullah University of Science and Technology, respectively, and both contain 40 sets of polarization

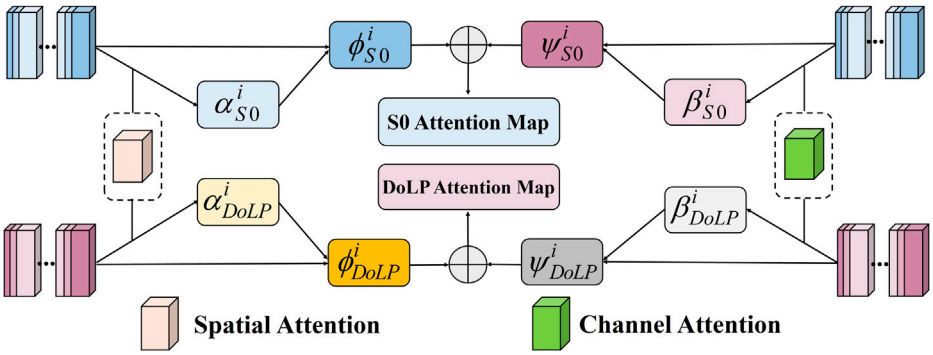


FIGURE 4
The framework of the AEU. α^i_k and β^i_k are obtained from SA and CA, respectively, while ϕ^i_k and ψ^i_k denote the enhanced features, k = S0, DoLP.

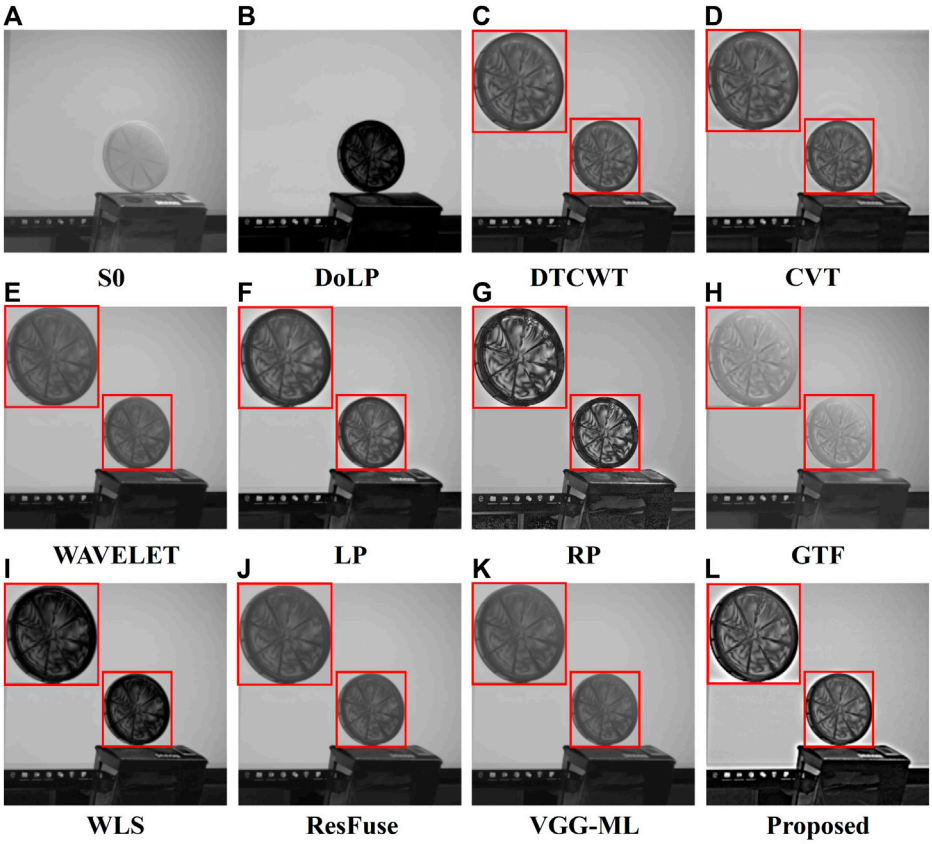


FIGURE 5
Qualitative fusion result of scene 1 on the public dataset. **(A)** S0; **(B)** DoLP; **(C)** DTCWT; **(D)** CVT; **(E)** WAVELET; **(F)** LP; **(G)** RP; **(H)** GTF; **(I)** WLS; **(J)** ResFuse; **(K)** VGG-ML; **(L)** Proposed. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

images. The experimental device is an Intel (R) Core (TM) i7-6700 CPU with 16 GB of RAM. The comparison algorithms include Dual-tree Complex Wavelet (DTCWT) [34], Curvelet Transform (CVT) [35], Wavelet [36], Laplacian Pyramid (LP) [37], Ration of Low-Pass Pyramid (RP) [38], Gradient Transfer Fusion (GTF) [39], WLS [40], ResFuse [11], and VGG-ML [16].

The evaluation metrics include information entropy (EN), spatial frequency (SF), standard deviation (SD), average gradient (AG), sum of difference correlation (SCD), and mutual information (MI). Among them, EN can reflect the information content of the fused image, SF reflects the rate of change of image grayscale, and both have a role in measuring

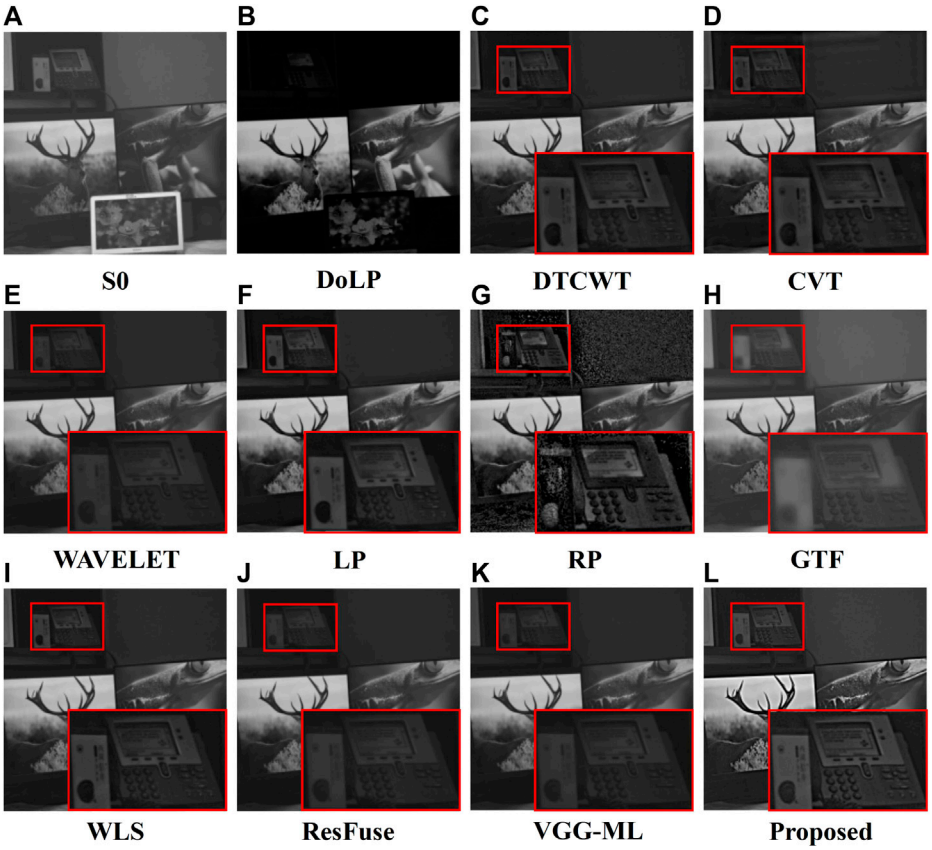


FIGURE 6
Qualitative fusion result of scene 2 on the public dataset. **(A)** S0; **(B)** DoLP; **(C)** DTCWT; **(D)** CVT; **(E)** WAVELET; **(F)** LP; **(G)** RP; **(H)** GTF; **(I)** WLS; **(J)** ResFuse; **(K)** VGG-ML; **(L)** Proposed. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

image quality with SD; AG can reflect the sharpness of the image; SCD can measure the information correlation between the fused image and the source image; and MI indicates the degree of correlation between images. Therefore, we use these metrics to comprehensively evaluate the fused image quality of different algorithms, and then verify the advantages of the proposed method.

4.2 Experimental results of the fusion algorithm

4.2.1 Results on the public dataset

The fusion results are shown in Figures 5, 6, where the key areas are marked using red boxes and enlarged. DTCWT and CVT are enhanced, but the texture details are less clear. Wavelet retains more balanced information, but the characterization effect for details is not sufficient. LP can retain the information from the source image, but the details are not clear enough. The fusion effect of RP is more outstanding, but there is a serious distortion problem, and the overall quality of the fused image is not ideal. The contrast of GTF is improved, but the focus is on retaining the information of S0, and the fused image is distorted and blurred. WLS has a fusion effect closer to the feature distribution of DoLP but retains little information from S0, and the overall contrast of the fused image is low. The fusion effect of both ResFuse and VGG-ML is relatively clear and natural, but the effect of these two algorithms on texture detail and overall contrast enhancement is still lacking. The comparison shows that our method can better balance the information of the source image, i.e., while retaining the high contrast and clear details of S0, it can also fully combine the polarization characteristics of DoLP.

The average calculation of the dataset images using the six metrics mentioned above and the experimental results are shown in Table 1. Our method achieves the best values in five metrics, EN, SF, SD, AG, and SCD, and the index values of SF and AG are improved by 12.963% and 40.152%, respectively, compared to the maximum values in the comparison algorithm. The best values of EN indicate that the fusion results of our method can obtain more information; the best values of SF and SD indicate that the fused images have higher quality; the metric values of AG are improved substantially, which represents a clearer, more detailed texture; and the best value of SCD can indicate the better fusion performance of the proposed network. In addition, the maximum value of MI in the metrics is obtained by VGG-ML. Although the MI value of our method is not optimal, the target fused image needs to have both

TABLE 1 Quantitative comparisons of the six metrics, i.e., EN, SF, SD, AG, SCD, and MI, on the public dataset. The best results are highlighted in bold.

Methods	EN	SF	SD	AG	SCD	MI
DTCWT	5.93668	4.05239	9.67117	2.50937	1.05129	3.32856
CVT	6.03409	4.05189	9.69706	2.56217	1.05134	3.00594
WAVELET	5.60282	2.48074	9.50487	1.49367	1.05394	4.25609
LP	5.96029	3.13571	8.73245	2.05446	0.74964	2.98191
RP	6.19134	6.65251	8.74362	2.78993	0.68997	2.42422
GTF	5.75619	2.85640	10.76809	1.75516	0.80165	4.06385
WLS	5.74227	4.24776	8.63091	2.68236	0.85460	3.56277
ResFuse	6.01416	2.59245	9.59885	1.57764	1.04945	3.70814
VGG-ML	5.64989	2.64256	9.53368	1.61344	1.06090	**4.29472**
Proposed	**6.55868**	**7.51489**	**11.06179**	**3.91015**	**1.17152**	3.84981

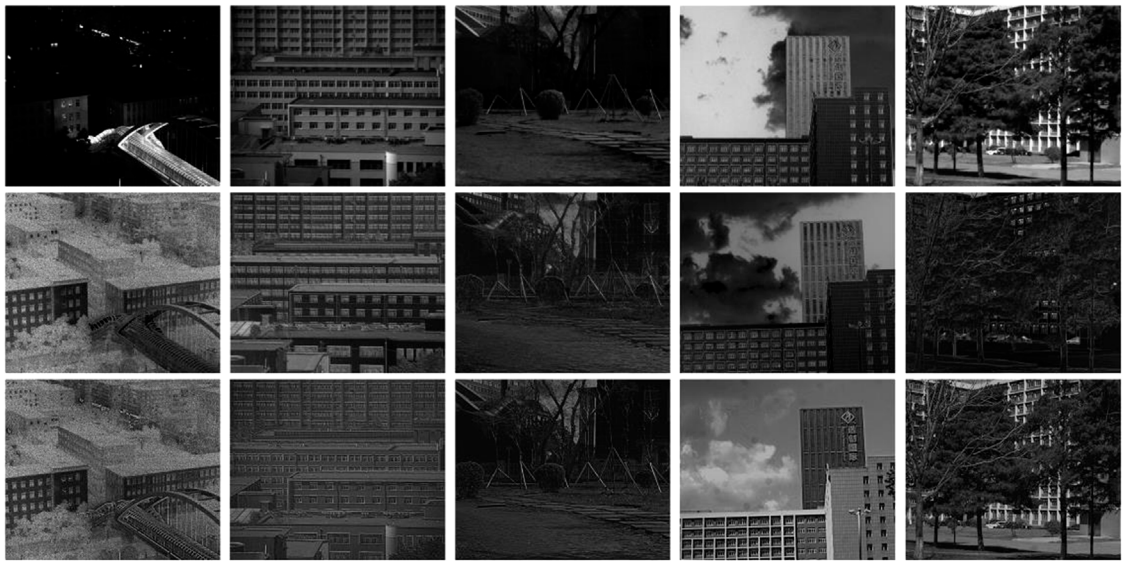


FIGURE 7
Our polarization dataset and fusion results. Row 1 contains five S0 images, Row 2 contains five DoLP images, and Row 3 contains five fused images obtained by the proposed method.

high contrast and texture, while other algorithms do not have these features at the same time. In a comprehensive comparison, our method can highlight the edge contours of the target more effectively and is more advantageous in enhancing the image contrast and details.

4.2.2 Results on our dataset

We use a focal plane polarization camera with a Sony IMX250MZR CMOS to photograph the campus scene and construct our dataset. Partial images and fusion results are shown in Figure 7. Our dataset mainly includes buildings, trees, etc. In outdoor scenes, which is quite different from the public dataset.

Two scenes in our polarization dataset are selected and compared with the above nine algorithms. As shown in Figures 8, 9, our method has a more obvious enhancement effect for both man-made and natural objects. The green and red boxes in Figure 8 show the fusion effect on the near and far views of the building, and our method has sharper details and higher contrast than the other methods. The green and red boxes in Figure 9 show the fusion effect on natural plants, and our method also has a more natural and visual perception.

As shown in Figure 10 fused images were selected in our dataset, and line graphs were drawn using the metrics mentioned above. The abscissa of Figures 10A–F represents the image sequence, and the ordinate represents the specific value of each metric. It can be found

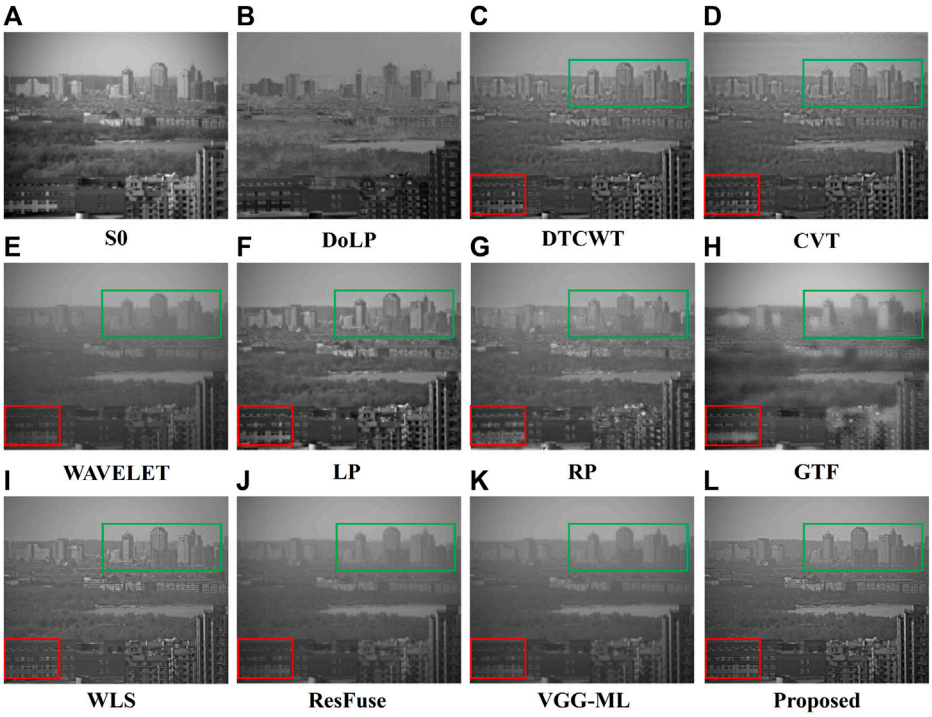


FIGURE 8
Qualitative fusion result of scene 1 on our dataset. **(A)** S0; **(B)** DoLP; **(C)** DTCWT; **(D)** CVT; **(E)** WAVELET; **(F)** LP; **(G)** RP; **(H)** GTF; **(I)** WLS; **(J)** ResFuse; **(K)** VGG-ML; **(L)** Proposed.

that our method achieves the best values for EN, SF, SD, and AG, thus verifying the outstanding advantages of our proposed method over other algorithms.

4.3 Ablation experiments

4.3.1 Ablation experiment of the QAU

Different fusion strategies are applied to the base layers without the use of an attention enhancement unit. First, the two commonly used strategies are compared; then, the fusion effect of the local energy weighting strategy is verified; and finally, EN and NR-IQA are added for experiments. Specifically, it includes: E1: average weighting strategy; E2: absolute maximum selection strategy; E3: local energy weighting strategy; and E4: QAU.

Qualitative comparisons are shown in Figure 11, where the focus areas are marked and enlarged using green and red boxes. The fusion effect of the average weighting strategy is more balanced, but the retention effect of details is not sufficient. The fused image obtained by using the absolute maximum selection strategy can retain the features of S0 better, but the information retention effect of DoLP is not ideal and does not balance the source image information reasonably. The fused images obtained by the local energy weighting strategy have more details, but the effect is still not outstanding. The fused images obtained by using the QAU can more fully reflect the different advantageous features of S0 and DoLP, while the overall effect is more natural.

As shown in Table 2, the maximum value of EN is obtained by the local energy weighting strategy, and the optimal value of SCD is obtained by the average weighting strategy. Although our method does not achieve optimal values for these two metrics, it achieves maximum values for SF, SD, AG, and MI. These metrics can objectively reflect the polarization image fusion effect and then verify the advantages of QAU compared with other fusion rules.

4.3.2 Ablation experiment of the VGG19

Experiments were conducted using ResNet50, ResNet101, ResNet152, VGG16, and VGG19 pre-trained on the MSCOCO dataset [41], respectively, while keeping other conditions constant, and quantitative metrics were calculated.

The experimental results are shown in Table 3. The VGG series networks have better fusion results than the residual series networks in our method, and VGG19 has higher metric values than VGG16 in all metrics. Among them, VGG19 achieves the highest metric values in EN, SF, SD, and AG, while SCD and MI have the best metric values in ResNet50. Therefore, after comparing the metric values, we selected VGG19 as the feature extraction network for the detail layers.

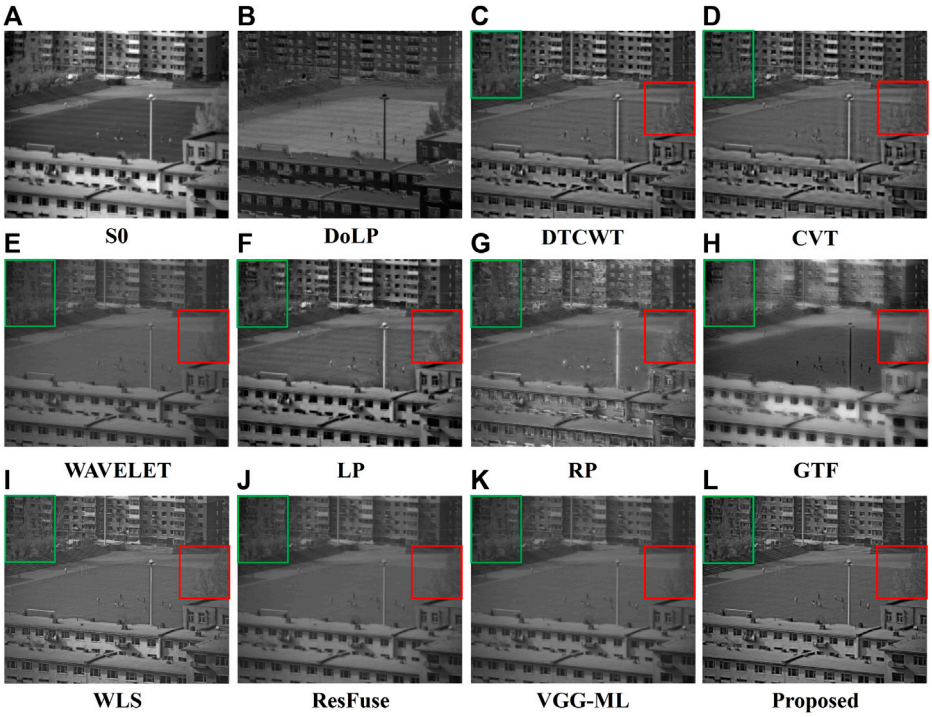


FIGURE 9
Qualitative fusion result of scene 2 on our dataset. **(A)** S0; **(B)** DoLP; **(C)** DTCWT; **(D)** CVT; **(E)** WAVELET; **(F)** LP; **(G)** RP; **(H)** GTF; **(I)** WLS; **(J)** ResFuse; **(K)** VGG-ML; **(L)** Proposed.

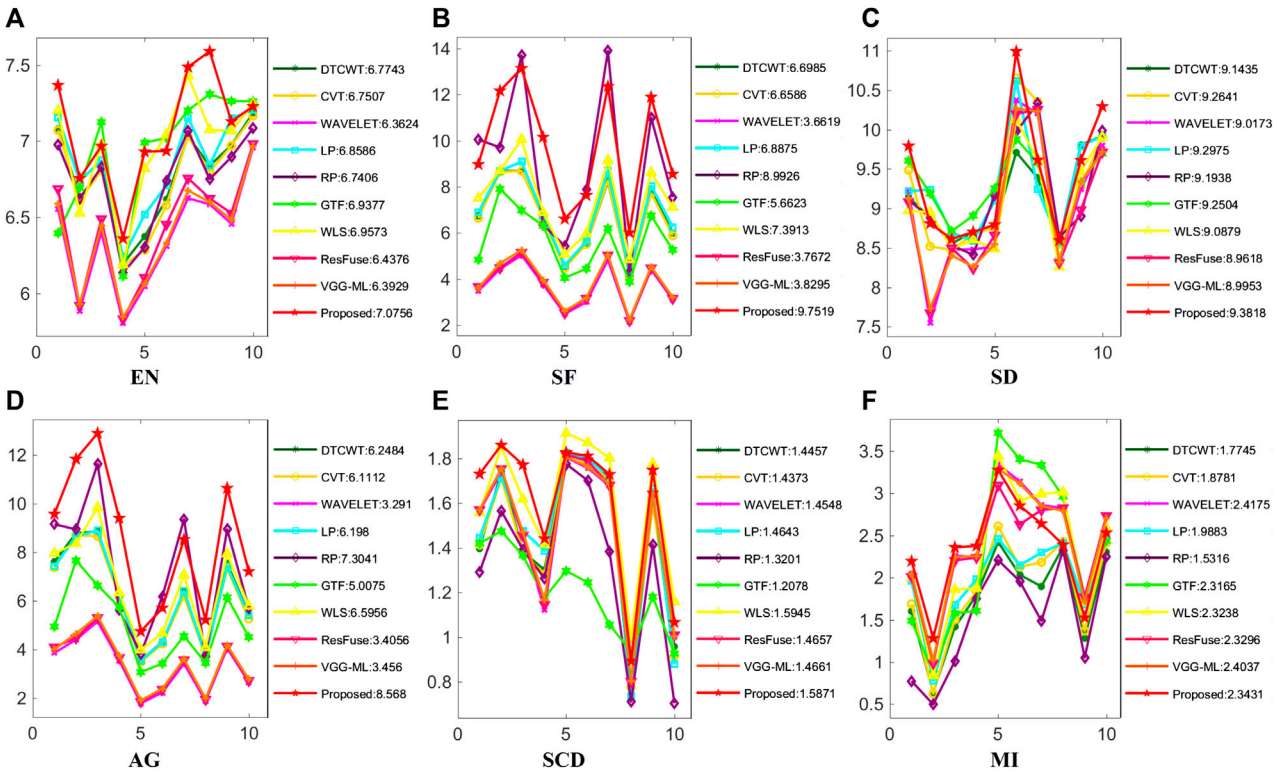


FIGURE 10
Quantitative comparison of six metrics in our dataset. **(A)** EN; **(B)** SF; **(C)** SD; **(D)** AG; **(E)** SCD; **(F)** MI.

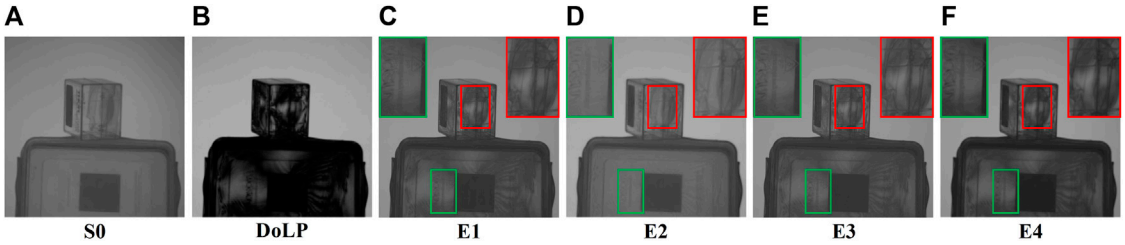


FIGURE 11
Qualitative fusion result of the QAU ablation experiment. **(A)** S0; **(B)** DoLP; **(C)** E1; **(D)** E2; **(E)** E3; **(F)** E4. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

TABLE 2 Quantitative comparison of the QAU ablation experiment. The best results are highlighted in bold.

Methods	EN	SF	SD	AG	SCD	MI
E1	5.63052	2.50456	9.51918	1.54976	**1.05928**	4.27311
E2	5.69372	2.86766	10.71943	1.76710	0.85805	4.28513
E3	**5.83124**	2.69043	9.63691	1.64626	0.94455	4.21881
E4	5.82460	**2.89027**	**10.90021**	**1.78798**	1.03466	**4.29613**

TABLE 3 Quantitative comparison of fusion results from different CNNs. The best results are highlighted in bold.

Methods	EN	SF	SD	AG	SCD	MI
ResNet50	6.15019	6.30726	11.05501	3.81852	**1.20510**	**3.57606**
ResNet101	6.14998	6.31661	11.05548	3.82214	1.20248	3.57515
ResNet152	6.15012	6.31411	11.05543	3.82139	1.20322	3.57527
VGG16	6.15518	6.49264	11.06048	3.89827	1.17482	3.55340
VGG19	**6.15587**	**6.51489**	**11.06179**	**3.91016**	1.17752	3.55981

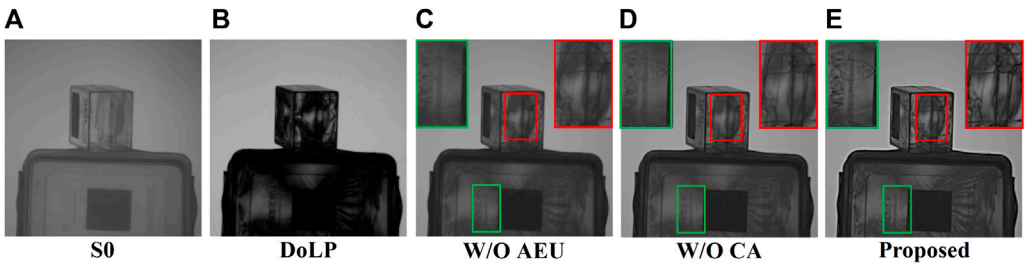


FIGURE 12
Qualitative fusion result of the AEU ablation experiment. **(A)** S0; **(B)** DoLP; **(C)** without the AEU; **(D)** without the CA; **(E)** Proposed. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

4.3.3 Ablation experiment of the AEU

To verify the effectiveness of the AEU, the following experiments were conducted separately. First, without AEU; then, CA is added; finally, both CA and SA are used, i.e., the proposed method. Qualitative comparisons are shown in Figure 12, and some areas are marked and enlarged using green and red boxes for better observation of the experimental effect.

When without AEU, the detail information of the fused image is not prominent; when only CA is used, the texture detail and overall contrast are somewhat improved, indicating

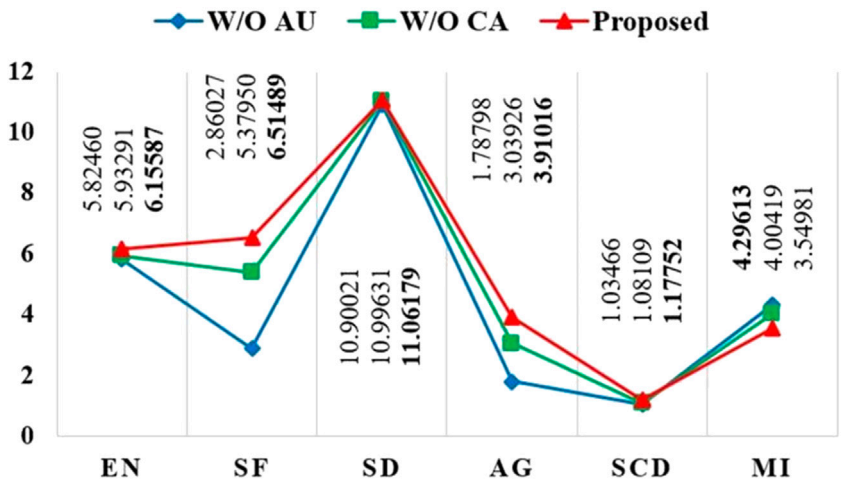


FIGURE 13
Quantitative comparison of the AEU ablation experiment. The proposed method is highlighted in red. Adapted from Linear polarization demosaicking for monochrome and colour polarization focal plane arrays by Qiu S et al., licensed under CC BY-NC 4.0 [33].

that the network can combine more information from S0 and DoLP at this time; and for the fusion result of using both CA and SA, it has a clearer texture and a more natural visual effect.

The quantitative comparison is shown in Figure 13. Compared with the fused images without AEU and with SA removed, our method has significant improvements in multiple metrics, which shows that AEU is beneficial to obtain higher-quality fused images.

5 Conclusion

This paper presents a dual-weighted polarization image fusion method that fuses S0 and DoLP to obtain a fused image of the target with high contrast and clear details at the same time. The source images are decomposed into base layers and detail layers, and the corresponding fusion strategies are designed based on quality assessment and attention mechanisms, respectively. A quality assessment unit is constructed for the fusion process of the base layers to ensure the high contrast of the fused image; a pre-trained VGG19 is used to extract the depth features of the detail layers, and a combined spatial and channel attention enhancement unit is constructed to achieve fuller preservation of scene information and texture contours to ensure the clear details and global information of the fused image. Experimental results on both public and our datasets show that the proposed method has more obvious enhancement effects in terms of contrast and detail texture for both small scene targets and complex outdoor environments, with better subjective visual perception and higher objective evaluation metrics compared to other algorithms. In future research work, we will explore how to reduce the complexity of the model while maintaining high fusion performance and combine the angle of polarization (AoP) to achieve better fusion 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 author.

Author contributions

JD and HZ responsible for paper scheme design, experiment and paper writing. JD and JL guide the paper scheme design and revision. MG, CC, and GC guide to do experiments and write papers. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the National Natural Science Foundation of China (62127813), the Project of Industrial Technology Research and Development in Jilin Province (2023C031-3), and the National Natural Science Foundation of Chongqing, China (cstc2021jcyj-msxmX0145).

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

References

1. Liu Y, Yang X, Zhang R, Albertini MK, Celik T, Jeon G. Entropy-based image fusion with joint sparse representation and rolling guidance filter. *Entropy* (2020) 22:118. doi:10.3390/e22010118

2. Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: A survey. *Inf Fusion* (2019) 45:153–78. doi:10.1016/j.inffus.2018.02.004

3. Chen J, Li X, Luo L, Mei X, Ma J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf Sci* (2020) 508:64–78. doi:10.1016/j.ins.2019.08.066

4. Li G, Lin Y, Qu X. An infrared and visible image fusion method based on multi-scale transformation and norm optimization[J]. *Inf Fusion* (2021) 71:109–129. doi:10.1016/j.inffus.2021.02.008

5. Li Y, Sun Y, Huang X, Qi G, Zheng M, Zhu Z. An image fusion method based on sparse representation and sum Modified-Laplacian in NSCT domain. *Entropy* (2018) 20:522. doi:10.3390/e20070522

6. Wang C, Wu Y, Yi Y, Zhao J. Joint patch clustering-based adaptive dictionary and sparse representation for multi-modality image fusion. *Machine Vis Appl* (2022) 33:69. doi:10.1007/s00138-022-01322-w

7. Gai D, Shen X, Chen H, Xie Z, Su P. Medical image fusion using the PCNN based on IQPSO in NSST domain. *IET Image Process* (2020) 14:1870–80. doi:10.1049/iet-ipr.2020.0040

8. Panigrahy C, Seal A, Mahato N. Parameter adaptive unit-linking dual-channel PCNN based infrared and visible image fusion. *Neurocomputing* (2022) 514:21–38. doi:10.1016/j.neucom.2022.09.157

9. Zhang J, Shao J, Chen J, Yang D, Liang B. Polarization image fusion with self-learned fusion strategy. *Pattern Recognition* (2021) 118:108045. doi:10.1016/j.patcog.2021.108045

10. Liu J, Duan J, Hao Y, Chen G, Zhang H. Semantic-guided polarization image fusion method based on a dual-discriminator GAN. *Opt Express* (2022) 24:43601–21. doi:10.1364/oe.472214

11. Li H, Wu X, Durrani T. Infrared and visible image fusion with ResNet and zero-phase component analysis. *Infrared Phys Technol* (2019) 102:103039. doi:10.1016/j.infrared.2019.103039

12. Li Y, Wang J, Miao Z, Wang J. Unsupervised densely attention network for infrared and visible image fusion. *Multimedia Tools Appl* (2020) 79:34685–96. doi:10.1007/s11042-020-09301-x

13. Wang S, Meng J, Zhou Y, Hu Q, Wang Z, Lyu J. Polarization image fusion algorithm using NSCT and CNN. *J Russ Laser Res* (2021) 42:443–52. doi:10.1007/s10946-021-09981-2

14. Zhu Z, Yin H, 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

15. Zhu Z, Zheng M, Qi G, 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

16. Li H, Li X, Kittler J. Infrared and visible image fusion using a deep learning framework. *2018 24th international conference on pattern recognition*. Beijing, China: ICPR (2018). pp. 2705–10. doi:10.1109/ICPR.2018.8546006

17. Liu X, Wang L. Infrared polarization and intensity image fusion method based on multi-decomposition LatLRR. *Infrared Phys Technol* (2022) 123:104129. doi:10.1016/j.infrared.2022.104129

18. Hu P, Wang C, Li D, Zhao X. An improved hybrid multiscale fusion algorithm based on NSST for infrared–visible images. *Vis Comput* (2023) 1–15. doi:10.1007/s00371-023-02844-8

19. Zou D, Yang B, Li Y, Zhang X, Pang L. Visible and NIR image fusion based on multiscale gradient guided edge-smoothing model and local gradient weight. *IEEE Sensors J* (2023) 23:2783–93. doi:10.1109/jsen.2022.3232150

20. Gao L, Guo Z, Zhang H, Xu X, Shen H. Video captioning with attention-based lstm and semantic consistency. *IEEE Trans Multimedia* (2017) 19:2045–55. doi:10.1109/tmm.2017.2729019

21. Brauwers G, Frasincar F. A general survey on attention mechanisms in deep learning. *IEEE Trans Knowledge Data Eng* (2021) 35:3279–98. doi:10.1109/tkde.2021.3126456

22. Wang X, Hua Z, Li J. Cross-UNet: dual-branch infrared and visible image fusion framework based on cross-convolution and attention mechanism. *Vis Comput* (2022) 1–18. doi:10.1007/s00371-022-02628-6

23. Li J, Li B, Jiang Y, Cai W. MSAt-GAN: A generative adversarial network based on multi-scale and deep attention mechanism for infrared and visible light image fusion. *Complex Intell Syst* (2022) 8:4753–81. doi:10.1007/s40747-022-00722-9

24. Liu Y, Zhou D, Nie R, Ding Z, Guo Y, Ruan X, et al. TSE_Fuse: Two stage enhancement method using attention mechanism and feature-linking model for infrared and visible image fusion. *Digital Signal Process.* (2022) 123:103387. doi:10.1016/j.dsp.2022.103387

25. Zhang S, Huang F, Liu B, Li G, Chen Y, Chen Y, et al. A multi-modal image fusion framework based on guided filter and sparse representation. *Opt Lasers Eng* (2021) 137:106354. doi:10.1016/j.optlaseng.2020.106354

26. Cao H, Luo X, Peng Y, Xie T. MANet: A network architecture for remote sensing spatiotemporal fusion based on multiscale and attention mechanisms. *Remote Sensing* (2022) 14:4600. doi:10.3390/rs14184600

27. Wang J, Xi X, Li D, Li F, Zhang G. GRPAFusion: A gradient residual and pyramid attention-based multiscale network for multimodal image fusion. *Entropy* (2023) 25:169. doi:10.3390/e25010169

28. Liu L, Liu B, Huang H, Bovik A. No-reference image quality assessment based on spatial and spectral entropies. *Signal Process. Image Commun* (2014) 29:856–63. doi:10.1016/j.image.2014.06.006

29. Bosse S, Maniry D, Müller K, Wiegand T, Samek W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans Image Process* (2018) 27:206–19. doi:10.1109/tip.2017.2760518

30. Hu S, Meng J, Zhang P, Miao L, Wang S. A polarization image fusion approach using local energy and MDLatLRR algorithm. *J Russ Laser Res* (2022) 43:715–24. doi:10.1007/s10946-022-10099-2

31. Xu H, Ma J, Le Z, Jiang J, Guo X. FusionDN: a unified densely connected network for image fusion. in Proceedings of the AAAI Conference on Artificial Intelligence, (2020). pp. 12484–91. doi:10.1609/aaai.v34i07.6936

32. Morimatsu M, Monno Y, Tanaka M, Okutomi M. Monochrome and color polarization demosaicking using edge-aware residual interpolation, in *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, (2020) pp. 2571–75. doi:10.1109/ICIP40778.2020.9191085

33. Qiu S, Fu Q, Wang C, Heidrich W. Linear polarization demosaicking for monochrome and colour polarization focal plane arrays. *Comput Graphics Forum* (2021) 40:77–89. doi:10.1111/cgf.14204

34. Lewis J, O'Callaghan R, Nikolov S, Bull D, Canagarajah R. Pixel-and region-based image fusion with complex wavelets. *Inf Fusion* (2007) 8:119–30. doi:10.1016/j.inffus.2005.09.006

35. Nencini F, Garzelli A, Baronti S, Alparone L. Remote sensing image fusion using the curvelet transform. *Inf Fusion* (2007) 8:143–56. doi:10.1016/j.inffus.2006.02.001

36. Chipman L, Orr T, Graham L. Wavelets and image fusion. Proceedings., International Conference on Image Processing, Washington, DC, USA. (1995). pp. 248–51. doi:10.1109/ICIP.1995.537627

37. Burt P, Adelson E. The laplacian pyramid as a compact image code. *IEEE Trans Commun* (1983) 31:532–40. doi:10.1109/tcom.1983.1095851

38. Toet A. Image fusion by a ratio of low-pass pyramid. *Pattern Recognition Lett* (1989) 9:245–53. doi:10.1016/0167-8655(89)90003-2

39. Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf Fusion* (2016) 31:100–9. doi:10.1016/j.inffus.2016.02.001

40. Ma J, Zhou Z, Wang B, Zong H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys Technol* (2017) 82:8–17. doi:10.1016/j.infrared.2017.02.005

41. Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context, in European conference on computer vision (2014) 8693: pp. 740–755. doi:10.1007/978-3-319-10602-1_48

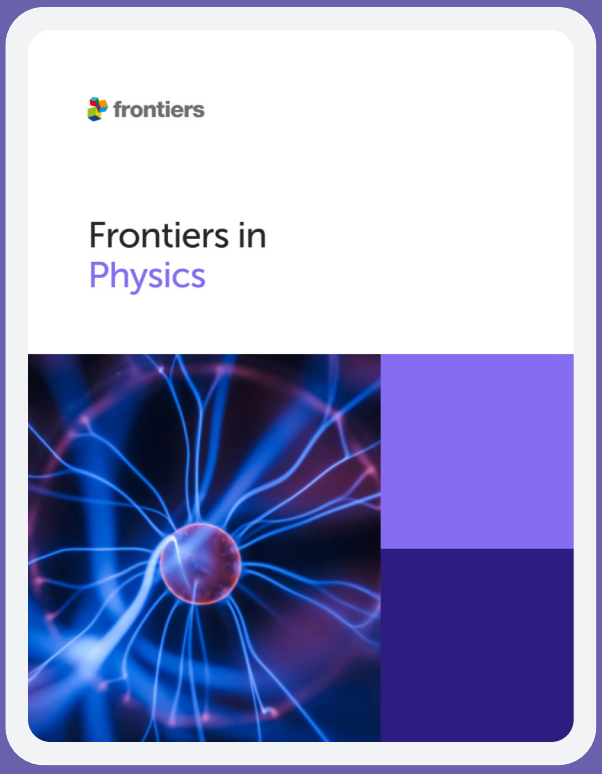
Frontiers in Physics

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

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact