RECENT ADVANCES IN ARTIFICIAL NEURAL NETWORKS AND EMBEDDED SYSTEMS FOR MULTI-SOURCE IMAGE FUSION

EDITED BY: Xin Jin, Jingyu Hou, Shin-Jye Lee and Dongming Zhou

PUBLISHED IN: Frontiers in Neurorobotics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-945-2
DOI 10.3389/978-2-88976-945-2

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

RECENT ADVANCES IN ARTIFICIAL NEURAL NETWORKS AND EMBEDDED SYSTEMS FOR MULTI-SOURCE IMAGE FUSION

Topic Editors:
Xin Jin, Yunnan University, China
Jingyu Hou, Deakin University, Australia
Shin-Jye Lee, National Chiao Tung University, Taiwan
Dongming Zhou, Yunnan University, China

Citation: Jin, X., Hou, J., Lee, S.-J., Zhou, D., eds. (2022). Recent Advances in Artificial Neural Networks and Embedded Systems for Multi-Source Image Fusion. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-945-2

Table of Contents

frontiers | Frontiers in Neurorobotics

TYPE Editorial
PUBLISHED 03 August 2022
DOI 10.3389/fnbot.2022.962170

OPEN ACCESS

EDITED AND REVIEWED BY
Jae-Ho Han,
Korea University, South Korea

*CORRESPONDENCE
Xin Jin
xinxin_jin@163.com

RECEIVED 06 June 2022
ACCEPTED 20 June 2022
PUBLISHED 03 August 2022

CITATION
Jin X, Hou J, Lee S-J and Zhou D (2022) Editorial: Recent advances in artificial neural networks and embedded systems for multi-source image fusion.
Front. Neurorobot. 16:962170.
doi: 10.3389/fnbot.2022.962170

Editorial: Recent advances in artificial neural networks and embedded systems for multi-source image fusion

Xin Jin[1]*, Jingyu Hou[2], Shin-Jye Lee[3] and Dongming Zhou[1]

[1]School of Software, Yunnan University, Kunming, China, [2]Institute of Technology Management, National Chiao Tung University, Hsinchu, Taiwan, [3]School of Information Technology, Deakin University, Geelong, VIC, Australia

KEYWORDS

artificial neural networks, embedded learning system, feature extraction, fusion strategy, image fusion

Editorial on the Research Topic
Recent advances in artificial neural networks and embedded systems for multi-source image fusion

Multi-source image fusion can help robotic systems to perceive the real world by fusing multi-source images from multiple sensors into a synthesized image that provides either a comprehensive or reliable description (Geng et al., 2016; Jin et al., 2017; Ma et al., 2017; Liu et al., 2018; Zhu et al., 2018; Zhang et al., 2021). At present, a large number of brain-inspired algorithm methods (or models) are aggressively proposed to accomplish image fusion asksk, and the artificial neural network has become one of the most popular techniques in the field of multi-source image fusion, especially deep convolutional neural networks (Liu et al., 2018; Jin et al., 2021). This is an exciting research field for the research community surrounding image fusion, with deep few-shot learning, unsupervised learning, application of embodied neural systems, and industrial applications.

How to develop a sound biological neural network and embedded system to fuse the multiple features of source images are two key questions that need to be addressed in the field of multi-source image fusion (Liu et al., 2019; Xu and Ma, 2021; Tang et al., 2022). Hence, studies of image fusion can be divided into two areas: first, new end-to-end neural network models for merging constituent parts during the image fusion process; second, the embodiment of artificial neural networks for image fusion systems. In addition, current booming techniques, including deep neural systems and embodied artificial intelligence systems, have been considered potential future trends for reinforcing the performance of image fusion.

In the first work entitled "Multi-Focus Color Image Fusion Based on Quaternion Multi-Scale Singular Value Decomposition (QMSVD)", Wan et al. employed multichannel quaternion multi-scale singular value to decompose the multi-focus color images, and a set of low-frequency and high-frequency sub-images was obtained. The activity and matching levels are exploited in the focus decision mapping of the low-frequency sub-image fusion, and a local contrast fusion rule based on the integration of high-frequency and low-frequency regions was also proposed. The fused images were finally reconstructed by inverse QMSVD. Experiments revealed that the color image fusion method has competitive visual effects.

The visual quality of images is seriously affected by bad weather conditions, especially on foggy days. To remove the fog in the image, Liu et al. introduced a method entitled "Single Image Defogging Method Based on Image Patch Decomposition and Multi-Exposure Image Fusion". In this method, the authors propose a single image defogging method based on image patch decomposition and multi-exposure fusion, which did not use any a priori knowledge of the scene depth information. First, a single foggy image was processed to produce a set of underexposed images, and then the underexposed and original images were enhanced and fused by guided filter and patch operation.

To protect the Tujia brocades that form part of the intangible cultural heritage, Shuqi He introduce a method using an unsupervised clustering algorithm for Tujia brocades segmentation, and a K auto-selection based on information fusion was also used. In this method, the cluster number K was calculated by fusing local binary patterns and gray-level co-occurrence matrix characteristic values. Thus, the clustering and segmentation operation can be performed on Tujia brocade images by adopting a Gaussian mixture model to get a rough preliminary segmentation image. Then, the voting optimization and conditional random filtering operation were used to optimize the preliminary segmentation and produce the final result.

In the fourth paper, Wu et al. propose fractional wavelet-based generative scattering networks (FrScatNets) in which fractional wavelet scattering networks are used as the encoder to extract image features, with deconvolutional neural networks acting as the decoder, to generate an image. Moreover, the authors also developed a feature-map fusion method to reduce the dimensionality of FrScatNet embeddings. In this work, the authors also discuss the application of image fusion in this study.

Conventional tensor decomposition is a kind of approximate decomposition model in which the image details may be lost in fused image reconstruction. To overcome this problem, Lu et al. introduced a work entitled "multi-modal image fusion based on matrix product state of tensor". In this work, source images were first separated into a third-order tensor, so that the tensor can be decomposed into a matrix product form by singular value decomposition, and then the Sigmoid function can be employed to fuse the key components. Thus, the fused image can be reconstructed by multiplying all the fused tensor components.

Lin et al. introduced an integrated circuit board object detection and image augmentation fusion model based on YOLO. In this paper, the authors first analyzed several popular region-based convolutional neural networks and YOLO models, and then they proposed a real-time image recognition model for integrated circuit board (ICB) in the manufacturing process. In this work, the authors first constructed an ICBs training dataset, and a preliminary image recognition model was then established to classify and predict ICBs. Finally, image augmentation fusion and optimization methods were used to improve the accuracy of the method.

Yu et al. report on a bottom-up visual saliency model in the wavelet domain. In this method, wavelet transform was first performed on the image to achieve four channels, and then discrete cosine transform was used to get the magnitude spectra and corresponding signum spectra. Third, wavelet decomposed multiscale magnitude spectra for every single channel were produced. Fourth, six multiscale conspicuity maps were generated for every single channel, and then the multiscale conspicuity maps of the four channels were fused. At last, a final saliency map after a scale-wise combination was obtained. The experimental results showed that the proposed model is effective.

Shi et al. propose an ensemble model for graph networks on imbalanced node classification, which uses GNNs as the base classifiers during boosting. In this method, the higher weights were set for the training samples that were not correctly classified by the previous classifiers. Besides, transfer learning was also employed to reduce computational cost and increase fitting ability. Experiments showed that the proposed method can achieve better performance than a graph convolutional network.

Deep neural networks have proven vulnerable to attack from adversarial examples. In response, Xie et al. propose a new noise data enhancement method, which only transforms adversarial perturbation to improve the transferability of adversarial examples with noise data enhancement and random erasing. Experiments have proved the effectiveness of this method.

The GAN-based method is difficult to converge completely to the distribution of face space in training. Yang et al. propose a face-swapping method based on a pretrained StyleGAN generator and designed a control strategy of the generator based on the idea of encoding and decoding to overcome the problem of GAN in this task. Experiments have shown that the performance of the proposed method is better than other state-of-the-art methods.

In the paper entitled "Adaptive fusion based method for imbalanced data classification", Liang et al. propose an ensemble method that combines data transformation and an adaptive weighted voting scheme for imbalanced data classification. They first utilized modified metric learning to obtain a feature space based on imbalanced data, and then the base classifiers were assigned different weights, adaptively. Experiments on multiple imbalanced datasets were performed to verify the performance of this algorithm.

In the work entitled "Multi-Exposure Image Fusion Algorithm Based on Improved Weight Function", Xu et al. proposed a multi-exposure image fusion method based on the Laplacian pyramid. Based on the Laplacian pyramid decomposition, an improved weight function was used to capture source image details. Six multi-exposure image fusion methods were compared with the proposed method on 20 sets of multi-exposure image sequences.

Sketch face recognition can match cross-modality facial images from sketch to photo, which is important in criminal investigations. Guo et al. introduced an effective cross task modality alignment network for sketch face recognition, and a meta learning training episode strategy was introduced to address the small sample problem. In this work, they propose a two-stream network to capture modality-specific and sharable features, and two cross task memory mechanisms to improve the performance of feature learning. At last, a cross task modality alignment loss is proposed to train the model.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was supported in part by the National Natural Science Foundation of China (Nos. 62101481 and 62002313), the Basic Research Project of Yunnan Province (Nos. 202201AU070033 and 202201AT070112), and the Key Laboratory in Software Engineering of Yunnan Province (No. 2020SE408).

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

Geng, P., Huang, M., Liu, S., Feng, J., and Bao, P. (2016). Multifocus image fusion method of Ripplet transform based on cycle spinning. *Multimedia Tools Applic.* 75, 10583–10593. doi: 10.1007/s11042-014-1942-1

Jin, X., Huang, S., Jiang, Q., Lee, S. J., Wu, L., and Yao, S. (2021). Semi-Supervised Remote Sensing Image Fusion Using Multi-Scale Conditional Generative Adversarial network with Siamese Structure. *IEEE J. Select Topics Appl. Earth Observ. Remote Sensing* 14, 7066–7084. doi: 10.1016/j.inffus.2017.10.007

Jin, X., Jiang, Q., Yao, S., Zhou, D., Nie, R., Hai, J., et al. (2017). A Survey of Infrared and Visual Image Fusion Methods. *Infrar. Phys. Technol.* 85, 478–501. doi: 10.1016/j.infrared.2017.07.010

Liu, S., Wang, J., Lu, Y., Hu, S., Ma, X., and Wu, Y. (2019). Multi-focus image fusion based on residual network in non-subsampled shearlet domain. *IEEE Access.* 7, 152043–152063. doi: 10.1109/ACCESS.2019.2947378

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., and Wang, X. (2018). Deep learning for pixel-level image fusion: Recent advances and future prospects. *Inf. Fusion* 42:158–173. doi: 10.1109/JSTARS.2021.3090958

Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., and Zhang, L. (2017). Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Trans. Image Proc.* 26, 2519–2532. doi: 10.1109/TIP.2017.2671921

Tang, L., Yuan, J., and Ma, J. (2022). Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* 82, 28–42. doi: 10.1016/j.inffus.2021.12.004

Xu, H., and Ma, J. (2021). EMFusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion* 76, 177–186. doi: 10.1016/j.inffus.2021.06.001

Zhang,. H., Xu,. H., and Tian, X. (2021). Image fusion meets deep learning: A survey and perspective. *Inf. Fusion* 76, 323–336. doi: 10.1016/j.inffus.2021.06.008

Zhu, P., Ding, L., Ma, X., and Huang, Z. (2018). Fusion of infrared polarization and intensity images based on improved toggle operator. *Optics Laser Technol.* 98, 139–151. doi: 10.1016/j.optlastec.2017.07.054

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fnbot.2021.695960

Multi-Focus Color Image Fusion Based on Quaternion Multi-Scale Singular Value Decomposition

Hui Wan[1,2], *Xianlun Tang*[3*], *Zhiqin Zhu*[3], *Bin Xiao*[1] *and Weisheng Li*[1]

[1] *College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China,* [2] *College of Computer and Information Science, Chongqing Normal University, Chongqing, China,* [3] *College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China*

Most existing multi-focus color image fusion methods based on multi-scale decomposition consider three color components separately during fusion, which leads to inherent color structures change, and causes tonal distortion and blur in the fusion results. In order to address these problems, a novel fusion algorithm based on the quaternion multi-scale singular value decomposition (QMSVD) is proposed in this paper. First, the multi-focus color images, which represented by quaternion, to be fused is decomposed by multichannel QMSVD, and the low-frequency sub-image represented by one channel and high-frequency sub-image represented by multiple channels are obtained. Second, the activity level and matching level are exploited in the focus decision mapping of the low-frequency sub-image fusion, with the former calculated by using local window energy and the latter measured by the color difference between color pixels expressed by a quaternion. Third, the fusion results of low-frequency coefficients are incorporated into the fusion of high-frequency sub-images, and a local contrast fusion rule based on the integration of high-frequency and low-frequency regions is proposed. Finally, the fused images are reconstructed employing inverse transform of the QMSVD. Simulation results show that image fusion using this method achieves great overall visual effects, with high resolution images, rich colors, and low information loss.

Keywords: multi-focus color image, image fusion, quaternion, singular value decomposition, multi-scale decomposition

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Shuaiqi Liu,
Hebei University, China
Bo Xiao,
Imperial College London, London, United Kingdom

****Correspondence:***
Xianlun Tang
tangxl_cq@163.com

Received: *15 April 2021*
Accepted: *17 May 2021*
Published: *23 June 2021*

Citation:
Wan H, Tang X, Zhu Z, Xiao B and Li W (2021) Multi-Focus Color Image Fusion Based on Quaternion Multi-Scale Singular Value Decomposition. Front. Neurorobot. 15:695960. doi: 10.3389/fnbot.2021.695960

INTRODUCTION

Image fusion is the process of combining the information from two or more images into a single image. It has been applied widely, ranging from medical analysis (Jin et al., 2018a,b, 2020), to remote sensing imaging and artificial fog removal (Zhu et al., 2020). An important branch of image fusion is multi-focus image fusion, which integrates images with different focal points into a full-focus image with global clarity and rich details. Multi-focus image fusion algorithms mainly include spatial domain methods, transform domain methods, and deep learning methods (Liu S. et al., 2020; Liu Y. et al., 2020).

The spatial domain methods can be grouped into pixel-based method, block-based method, and region-based method (Jin et al., 2018a,b; Qiu et al., 2019; Xiao et al., 2020). Compared with the pixel-based method, the other two use the spatial correlation of adjacent pixels to guide image fusion to avoid contrast reducing and detail loss in the fusion images. First, the original images are divided into a number of blocks or regions, and then the focus level and sharpness of each block or region is measured by image intensity information. Finally, a block or region with a higher degree of focus as part of the fusion image is selected. Vishal and Vinay (2018) proposed a block-based spatial domain multi-focus image fusion method, and used spatial frequency to measure the focus level of the blocks. Duan et al. (2018) proposed a segmentation scheme based on enhanced LSC, which embeds the depth information of pixels in the clustering algorithm for multi-focus image fusion. The main advantage of fusion methods based on spatial domain lies in the fact that simple to implement, it can obtain the focus measure with low computational complexity. However, the quality of image fusion is relevant to the selection of image block sizes or the segmentation algorithms. When the size of the image block is not properly selected, the fusion image may generate a "block effect." And if the segmentation algorithm fails to segment the region accurately, the focused region cannot be determined and extracted correctly.

In the transform domain approach, various multi-scale decomposition (MSD) methods are applied to multi-focus image fusion. Multi-scale decomposition algorithm mainly includes pyramid transform (Burt and Kolczynski, 1993; Du et al., 2016), wavelet transform (Gonzalo and Jesús, 2002; Jaroslav et al., 2002) and multi-scale geometric analysis (Li et al., 2017, 2018; Liu et al., 2019a). Compared with the pyramid and wavelet transforms, though the multi-scale geometric analysis method outperforms the pyramid and wavelet transforms in feature representation and excels in capturing multi-directional information and translation invariance, it is not time-efficient when it comes to decomposition and reconstruction. In addition to traditional multi-scale decomposition methods mentioned above, some other multi-scale fusion methods have been proposed. Zhou et al. (2014) proposed a novel image fusion scheme based on large and small dual-scale decomposition. In this scheme, the two-scale method is used to determine the image gradient weight, and removes the influence of anisotropic blur on the focused region detection effectively. An and Li (2019) introduced a novel adaptive image decomposition algorithm into the field of image processing, which can fast decompose images and has multi-scale characteristics. Zhang et al. (2017) proposed a multi-scale decomposition scheme by changing the size of the structural elements, and extracting the morphological gradient information of the image on different scales to achieve multi-focus image fusion. Ma et al. (2019) proposed a multi-focus image fusion method based on to estimate a focus map directly using small-scale and large-scale focus measures. Naidu (2011) proposed a novel method of multi-focus images fusion. In this method, multi-scale analysis and singular value decomposition are combined to perform multi-scale singular value decomposition on multi-focus images to obtain low-frequency sub-images and high-frequency sub-images of different scales. This multi-scale decomposition method has the stability and orthogonality of SVD. Since no convolution operation is required, the decomposition speed is fast.

Deep learning methods, which can be further grouped into classification model based methods and regression model based methods (Liu Y. et al., 2020). In the classification model, Liu et al. (2017) first introduced convolutional neural networks (CNN) into the field of multi-focus image fusion. With this method, the activity level measurement and the fusion rule can be jointly generated by learning a CNN model. In the regression model, Li et al. (2020) proposed a novel deep regression pair learning convolutional neural network for multi-focus image fusion. This method directly converts the entire image into a binary mask as the input of the network without dividing the input image into small patche, thereby solving the problem of the blur level estimation around the focused boundary due to patche division. These methods can extract more image features through self-learning of the deep network, and carry out image fusion based on these features. However, the difficulties in training a large number of parameters and large datasets have directly affected the image fusion efficiency and quality. Compared with deep learning methods, the conventional fusion methods are more extensible and repeatable, facilitating real-world applications. Thus, this paper mainly aims to improve the conventional multi-focus image fusion algorithms.

Most of the existing multi-focus image fusion algorithms mentioned above can process gray and color multi-focus images. As for the color multi-focus image fusion, each color channel is fused separately, and then combined to get the final fused image (Naidu, 2011; Liang and He, 2012; Aymaz and Köse, 2019). These traditional fusion methods ignore the inter-relationship between the color channels, which will lead to hue distortions and blur in the image fusion process. To solve the above problems, this paper proposes a novel mathematical model for color images based on quaternion matrix analysis. This model considers the human visual characteristics and interaction between pixels in color images and combines quaternion with multi-scale singular value decomposition (MSVD) (Kakarla and Ogunbona, 2001; Naidu, 2011). In this method, the three color components of a color image are decomposed as a whole to extract the rich color and detail information. Firstly, the three color components of the pixel are represented by three imaginary parts of a quaternion. Secondly, the multi-focus color image represented by the quaternion matrix is decomposed into a low-frequency sub-image and several high-frequency sub-images using multi-scale singular value decomposition (MSVD). The former contains the approximate structure and color information of the source image, the latter contains detailed features. Then, the low-frequency component and the high-frequency component are respectively fused based on different fusion rules. The designed fusion rule makes full use of the decomposition coefficient represented by the quaternion and applies the structural information and color information of the image to the fusion. Finally, the fusion components are used to reconstruct the fusion image. The fused image can more accurately maintain the spectral characteristics of the color channel. We define this method as quaternion multi-scale singular value decomposition (QMSVD). The main innovations of this method are listed below:

- The combination of quaternion and multi-scale singular value decomposition is applied to multi-focus color image fusion for the first time. That is, the color image represented by the quaternion is decomposed by multi-scale singular value decomposition, and the sub-images obtained by decomposition better retain the structure and color information of the original image.

- The multi-channel is introduced into the QMSVD for the first time, and achieve the purpose of extracting the salient features on the channels of different decomposition layers for image fusion.
- In the fusion of low-frequency sub-images, in order to make full use of the color information of the image, an improved fusion rule of local energy maximization is proposed, and the fusion rule introduces the color difference between pixels and combines local energy. In the fusion of high-frequency sub-images, the fusion results of low-frequency coefficients are incorporated into the fusion of high-frequency sub-images, and a local contrast fusion rule based on the integration of high-frequency and low-frequency regions is proposed.

The structure of this paper is organized as follows. Section Multi-Scale Singular Value Decomposition of a Color Image introduces the concept of multi-scale singular value decomposition of a color image. Section Multi-Focus Color Image Fusion Based on QMSVD proposes multi-focus color image fusion model based on QMSVD. Section Experimental Results and Discussion we compare and analyze the results obtained through the state-of-the-art methods. Finally, conclusions for this paper are made in section Conclusion.

MULTI-SCALE SINGULAR VALUE DECOMPOSITION OF A COLOR IMAGE

To decompose the color image we integrate quaternion representation of color image with multi-scale decomposition. In this way, the approximate and detailed parts represented by quaternion can be obtained. The two parts are respectively fused, and the fused components are used to reconstruct the fusion image.

Quaternion Representation of a Color Image

Quaternions were discovered in 1843 by the Irish mathematician and physicist William Rowan Hamilton. It is extension of ordinary complex number, which extends ordinary complex numbers from a two-dimensional space to a four-dimensional space. A quaternion is composed of a real part and three imaginary parts. The operations of the three imaginary parts are equivalent, which makes it very suitable for describing color images and expressing the internal connection of color channels. The three color channels of the image can be represented by three imaginary parts of quaternion (Chen et al., 2014; Xu et al., 2015; Grigoryan and Agaian, 2018). The general form of a quaternion is $q = q_a + q_b i + q_c j + q_d k$. It contains one real part q_a and three imaginary parts $q_b i$, $q_c j$ and $q_c k$, if the real part q_a of a quaternion q is zero, q is called a pure quaternion. The conjugation of quaternions is defined as:

$$q^* = q_a - q_b i - q_c j - q_d k \tag{1}$$

The modulus of a quaternion is defined as:

$$|q| = \sqrt{qq^*} = \sqrt{q_a^2 + q_b^2 + q_c^2 + q_d^2} \tag{2}$$

The rotation theory of quaternions is stated as follows:

In the three-dimensional space, u is a unit of pure quaternion, and the modulus is $|u| = 1$. If $R = e^{u\theta}$, then RXR^* indicates that the pure quaternion X is rotated by 2θ radians about the axis. u and θ are defined as:

$$u = \frac{1}{\sqrt{q_b^2 + q_c^2 + q_d^2}}(q_b i + q_c j + q_b i q_d k)$$

$$\theta = \begin{cases} \tan^{-1}\sqrt{q_b^2 + q_c^2 + q_d^2}/q_a, & q_a \neq 0 \\ \pi/2 & q_a = 0 \end{cases}$$

Let $u = (i + j + k)/\sqrt{3}$, which represents a three-dimensional grayscale line in RGB space. The three color components of the pixels on the grayscale line are all equal. Let $\theta = \pi/2$, that is:

$$RXR^* = e^{u\pi/2}X(e^{u\pi/2})^* = (i+j+k)/\sqrt{3}{}^*X^*(-i-j-k)/\sqrt{3} \tag{3}$$

Equation (3) means that X is rotated around the gray line u by 180 degree. That is, X is turned to the opposite direction with u as the axis of symmetry. Then, the pixel $X + RXR^*$ falls on the grayscale line.

A color image can be represented as a pure quaternion, that is:

$$f(x,y) = f_R(x,y) \cdot i + f_G(x,y) \cdot j + f_B(x,y) \cdot k \tag{4}$$

In Equation (4), $f_R(x,y), f_G(x,y), f_B(x,y)$ represent the R, G, and B color channel components of the color image, respectively. The x, y represent the rows and columns of the color image matrix, where the pixels reside. Such a color image can be represented by a quaternion matrix, and the processing of the color image can be performed directly on the quaternion matrix. In contrast with the traditional approaches, which convert a color image to a grayscale one or process each color channel separately, the quaternion method can process the color image as a whole.

Multi-Scale Decomposition of a Color Image

The singular value decomposition is an important matrix decomposition in linear algebra (Liu et al., 2019b), and it is to decompose the image matrix diagonally according to the size of the eigenvalues. There is no redundancy among the decomposed images, and it is suitable to use different fusion rules for the fusion of each sub-image. We extend decomposition to the multi-scale form in this section. Using multi-scale can perform image fusion in different scales and different directions.

X_q is the quaternion matrix form of the color image $f(x,y)$. The rank of the $m \times n$ quaternion matrix X_q is r. Given the $m \times m$ quaternion unitary matrix U_q and $n \times n$ quaternion unitary matrix V_q, we can get:

$$(U_q)^H X_q V_q = \begin{bmatrix} \Lambda_r & 0 \\ 0 & 0 \end{bmatrix} \equiv \Lambda \in R^{m \times n} \tag{5}$$

where the superscript H represents conjugate transpose, and $\Lambda_r = diag\{\lambda_1, \lambda_2, \cdots, \lambda_r\}, \lambda_i (1 \leq i \leq r)$ is the singular value of X_q, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. It follows that the singular value decomposition of the quaternion matrix X_q is:

$$X_q = U_q \begin{bmatrix} \Lambda_r & 0 \\ 0 & 0 \end{bmatrix} (V_q)^H \tag{6}$$

In Equation (6), $U_q(U_q)^H = I^{m \times m}$, $V_q(V_q)^H = I^{n \times n}$Unit matrix.

The multi-scale singular value decomposition of a color image represented by a quaternion can be realized, according to the ideas proposed in Naidu (2011). The $M \times N$ color image X_q, represented by the quaternion, is divided into non-overlapping $m \times n$ blocks, and each sub-block is arranged into an $mn \times 1$ vector. By combining these column vectors, a quaternion matrix $X_q{}'$ with a size of can be obtained. The singular value decomposition of $X_q{}'$ is:

$$X_q{}' = U_q{}' \Lambda' (V_q{}')^H \tag{7}$$

$U_q{}'$ and $V_q{}'$ are orthogonal matrices, and Λ' is a non-singular diagonal matrix after $X_q{}'$ decomposition. According to Equation (7):

$$S = (U_q{}')^H X_q{}' = \Lambda' (V_q{}')^H \tag{8}$$

the size of the quaternion matrix S is $mn \times MN/mn$.

According to the singular value decomposition mentioned above, the first column vector of $U_q{}'$ corresponds to the maximum singular value. When it is left multiplied by the matrix $X_q{}'$, the first row $S(1,:)$ of S carries the main information from the original image, which can be regarded as the approximate, or smooth component of the original image. Similarly, the other rows $S(2:mn,:)$ of S correspond to smaller singular values, which retain such detailed information as the texture and edge. Therefore, through singular value decomposition, the image can be decomposed into low-frequency and high-frequency sub-images by the singular value to achieve the multi-scale decomposition of the image. In the QMSVD approach, decomposition is goes layer by layer, repeating the process above. In repeated decomposition, the approximate component $S(1,:)$ of the upper layer is used to replace the next layer of X_q.

When the original image is divided into $m \times n$ blocks, according to the different values of m and n, QMSVD can be called $(m \times n)$-channel QMSVD. For example: when $m = 2$ and $n = 2$, it is called four-channel QMSVD when $m = 2$ and $n = 3$ or $m = 3$ and $n = 2$, it is called six-channel QMSVD, when $m = 2$ and $n = 4$ or $m = 4$ and $n = 2$, it is called eight-channel QMSVD.

We take six-channel QMSVD as an example to illustrate the decomposition structure of each layer. Let $m = 2$, $n = 3$, and $m \times n = 6$:

$$\phi_{LL} = S(1:) \begin{matrix} \psi_{H1} = S(2:), \psi_{H2} = S(3:), \psi_{H3} = S(4:) \\ \psi_{H4} = S(5:), \psi_{H5} = S(6:) \end{matrix} \tag{9}$$
$$X_q \rightarrow \{\phi_{LL}, \{\psi_{H1}, \psi_{H2}, \psi_{H3}, \psi_{H4}, \psi_{H5}\}, U\}$$

In Equation (9), the lowest-resolution approximation component vector is ϕ_{LL}, the detail component vectors are $\{\psi_{H1}, \psi_{H2}, \psi_{H3}, \psi_{H4}, \psi_{H5}\}$, and the eigenvector matrix is U. During the transformation of the lower layer, ϕ_{LL} is replaced with X_q, the decomposition operates by Equation (9) and the next layer decomposition is obtained, and the multilayer decomposition of the image can be obtained by repeating the process. Because the decomposition process is reversible, the original image can be reconstructed by inverse transformation of QMSVD.

The QMSVD method proposed in this paper, the MSVD (Naidu, 2011) method and the QSVD (Bihan and Sangwine, 2003) method all decompose the image through singular value decomposition, but they have their distinct characteristics. In Naidu (2011), the MSVD is mainly a decomposition method for gray images. When decomposing a color image, the MSVD method is used on each color channel, and then combine the three decomposed color channels to obtain a decomposed color image. This decomposition method of channel information separation ignores the correlation between channels and take no account of color information of image. The QMSVD method overcomes the shortcomings of the MSVD method, and can maintain the correlation between color channels while decomposing color images. Compared with the QMSVD method, QSVD directly decomposes color images to get the eigenvalues and corresponding eigenvectors. Then, according to experience, we use the truncation method on QSVD to divide the eigenvalues in a descending order into different segments to realize image decomposition. However, the decomposition process based on experience truncation method lacks a definite physical meaning. In order to ascribe a clear physical and geometric meaning to the decomposition process, the multi-channel QMSVD is introduced, which directly decomposes the image into low-frequency and high-frequency components of different scales according to the size of eigenvalues.

Figure 1 compares the results achieved by three decomposition methods. It can be seen that: (1) The QMSVD method directly decomposes the color image into a low-frequency component and three high-frequency components. The low frequency component is an approximation of the original image, which retains the characteristics of the original image in terms of structure and color. The high-frequency components extract the edge and contour features of the original image. (2) The MSVD method does not directly decompose the color image. First, decompose each color channel, and then combine the decomposed components into low-frequency components and high-frequency components. Compared with the QMSVD method, the low-frequency component does not retain the color characteristics of the original image. As it can be seen from the **Figure 1**, the main color of the low-frequency component is blue, while the main color of the original image is red. The high-frequency component extracts the edge and contour features of the original image, but does not have the fine features extracted by the QMSVD method. This is due to the fact that the edge features of each component cannot be completely overlapped when the components are combined. (3) Compared with the QMSVD method, the QSVD method is not strong on extracting detailed features. It can be seen from the **Figure 1** that the main structure and color information are in the decomposed image corresponding to the first feature value, and the other feature values are truncated into three segments, corresponding to the three decomposed images respectively,

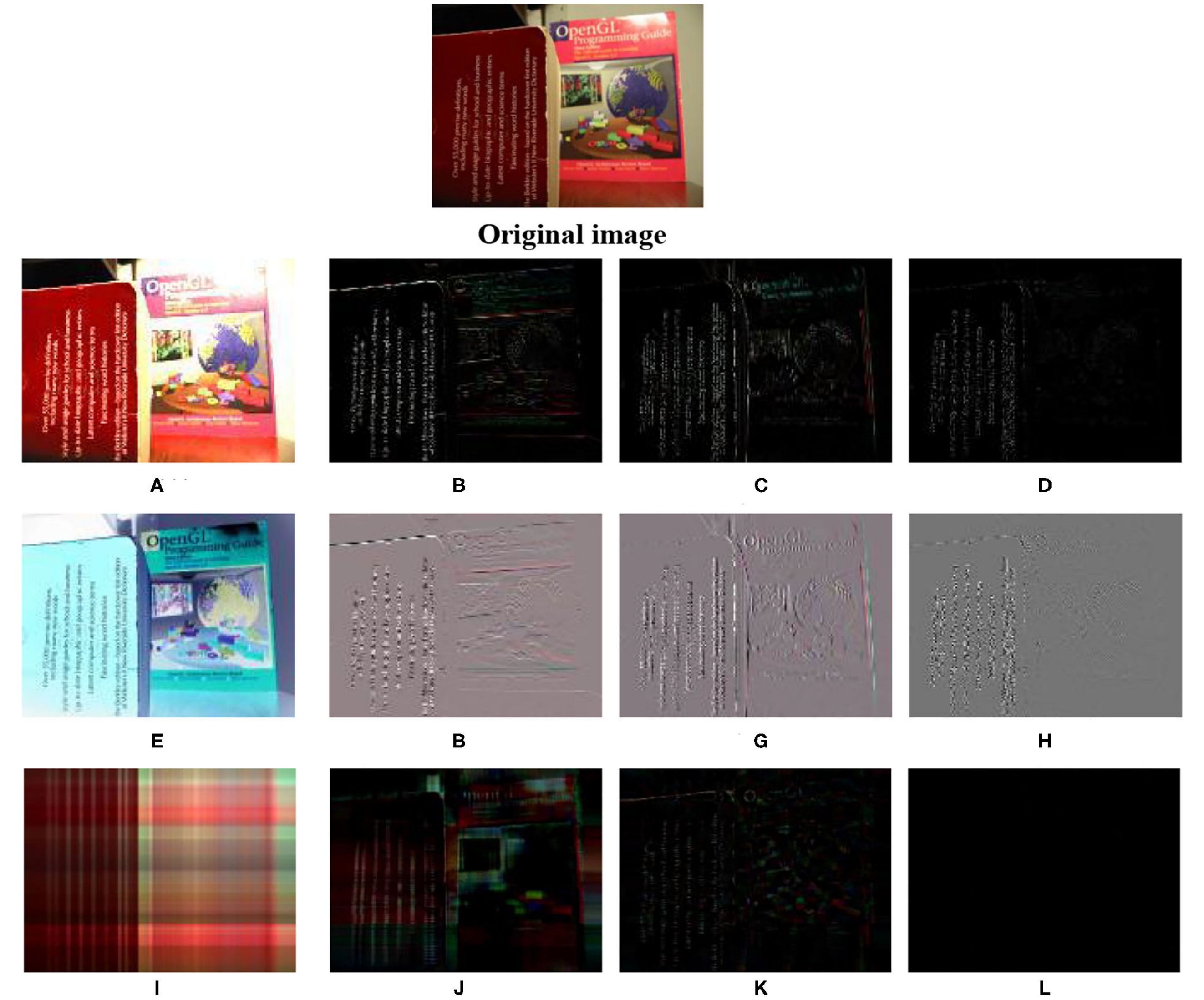


FIGURE 1 | This figure shows the decomposition of color image by QMSVD, MSVD, and QSVD. **(A)** The low-frequency image of the origianl image after decomposed by QMSVD, and **(B–D)** the high-frequency images of the origianl image. **(E)** The low-frequency image of the original image after decomposed by MSVD, and **(F–H)** the high-frequency images of the original image. **(I)** The decomposition image corresponding to the first eigenvalue of the original image decomposed by QSVD, and **(J)** the decomposition image corresponding to the eigenvalue truncated from the 2th to the 25th after QSVD decomposition, **(K)** the decomposition image corresponding to the eigenvalue truncated from the 26th to the 50th, **(L)** the decomposition image corresponding to the eigenvalue truncated from the 51th to the 240th. The eigenvalues are arranged from large to small.

and these images only carry a small amount of detailed features. Since the QSVD method is mainly used for image compression, in the experimental comparison part, we only compare QMSVD with MSVD methods.

MULTI-FOCUS COLOR IMAGE FUSION BASED ON QMSVD

Low-Frequency Component Fusion Rules

The low-frequency sub-image of QMSVD reflects the overall characteristics of the color original image. Commonly used low-frequency sub-image fusion rules include weighted average and maximum local energy. The weighted average rule is to get the fusion coefficient by weighted average of the low frequency coefficients in the same position of the images, which will result in the decline in the contrast of the fused image. The rule of maximum local energy is to compare the energy of low-frequency coefficients at the same position of the images, and choose the higher energy as the fusion coefficient. This fusion rule only considers the local energy of the image, and does not factor in the color information contained in the color image, so the visual effect of the color fusion image is not desirable. In order to overcome

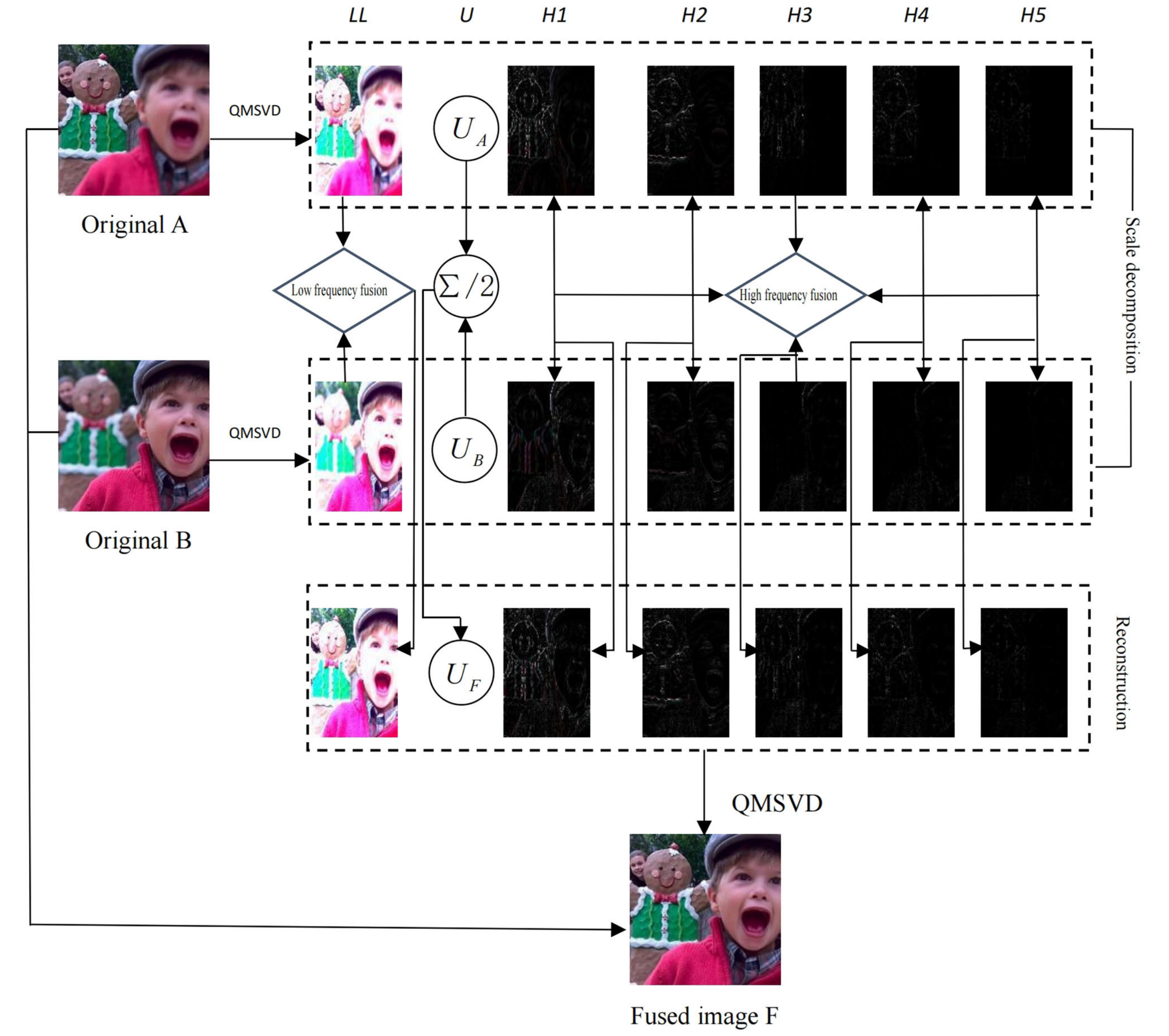


FIGURE 2 | Fusion of sub-images by QMSVD with six channels. *LL* is the low-frequency component of the decomposed image, *H1–H5* are the high-frequency components of the decomposed image, U_A and U_B are the orthogonal matrices of the decomposed image, and $U_F = (U_A + U_B)/2$.

the inadequacy, QMSVD uses a quaternion to represent the color image, and calculates the color differences between two color pixels based on the quaternion rotation theory. The coefficient window energy is used as activity level of the low frequency component, and the color difference between the color pixels in the center of the coefficient window is deemed as the matching level, with both jointly participating in the decision mapping.

Activity Level

Given the human visual system is sensitive to local variation, local window energy is used as the measurement of activity level. Local areas with larger variance exhibit greater contrast between pixels, and stronger window activity level. In contrast, pixel values more uniform in local areas with smaller variance, display weaker window activity level. Therefore, the pixel with the highest contrast in the low-frequency coefficient is selected as the fusion result.

$$d_S^j(x,y) = \left| C_S^j(x,y) - \underset{(x',y')\in p}{mean}\,(C_S^j(x+x',y+y')) \right| \qquad (10)$$

Where S represents the two color multi-focus images A and B to be fused, j represents the decomposition scale, $C_S^j(x,y)$ is the low-frequency sub-band coefficient of the original image S on scale j at pixel (x,y), P is the range of the coefficient window, $d_S^j(x,y)$ is the activity level of $C_S^j(x,y)$ at pixel (x,y), and $mean(\cdot)$ represents mean filtering. Experiments show that the visual effect after image fusion is the most optimal when P uses 3×3 local windows.

Matching Level

The matching level between A and B pixels of two color multi-focus images can be measured by the color differences between them, which can be calculated with the quaternion rotation theory (Jin et al., 2013). As the color difference includes chromaticity and luminance, the formula for calculating the matching level is as follows:

$$m_{AB}^j(x,y) = t\left|Q(q_1,q_2)\right| + (1-t)\left|I(q_1,q_2)\right| \tag{11}$$

In Equation (11), $q_1 = r_1 i + g_1 j + b_1 k$ and $q_2 = r_2 i + g_2 j + b_2 k$ are the pixels represented by quaternions in the color original images A and B, respectively. $Q(q_1,q_2)$ and $I(q_1,q_2)$ denote the differences in chromaticity and luminance, respectively, between q_1 and q_2, the weight $t \in [0,1]$ indicates the relative importance of chromaticity and luminance, and j represents the decomposition scale. According to the theory of quaternion rotation, the relationship between q_1 and q_2 can be expressed as $q_3 = q_1 + Rq_2R^* = r_3\cdot i + g_3\cdot j + b_3\cdot k, R = e^{u\pi/2}, u = (i+j+k)/\sqrt{3}$. If the chromaticity of q_1 is similar to that of q_2, q_3 should be near the grayscale line u, and the chromaticity difference between q_1 and q_2 can be expressed by the following equation:

$$\begin{aligned} Q(q_1,q_2) = {} & (r_3-(r_3+g_3+b_3)/3)\cdot i + (g_3-(r_3+g_3 \\ & +b_3)/3)\cdot j + (b_3-(r_3+g_3+b_3)/3)\cdot k \end{aligned} \tag{12}$$

When $Q(q_1,q_2)$ is small, the chromaticity of q_1 and q_2 are similar; when $Q(q_1,q_2)=0$, q_1 and q_2 have the same chromaticity. The difference in luminance between q_1 and q_2 can be illustrated as:

$$I(q_1,q_2) = (r_1-r_2)/3 + (g_1-g_2)/3 + (b_1-b_2)/3 \tag{13}$$

According to Equations (11–13), the size of $m_{q_1q_2}^j$ is proportional to the color difference between q_1 and q_2. Therefore, the matching level between the two pixels can be measured by the size of the color difference.

Decision Plan

The decision value of the color image focus judgment is determined by the activity level and matching level of the local window. They are obtained by Equations (10, 11), respectively. The decision value is calculated by the following formula:

$$d^j(x,y) = \begin{cases} 1, & if\ \ m_{AB}^j(x,y) > T\ and\ d_A^j(x,y) \geq d_B^j(x,y) \\ 0, & if\ \ m_{AB}^j(x,y) > T\ and\ d_A^j(x,y) < d_B^j(x,y) \\ \frac{1}{2}+\frac{1}{2}\left(\frac{1-T}{1-m_{AB}^j(x,y)}\right), & if\ \ m_{AB}^j(x,y) \leq T\ and\ d_A^j(x,y) \geq d_B^j(x,y) \\ \frac{1}{2}-\frac{1}{2}\left(\frac{1-T}{1-m_{AB}^j(x,y)}\right), & otherwise \end{cases} \tag{14}$$

According to the decision value $d^j(x,y)$, the fused low-frequency image can be obtained using $F_L^j(x,y) = d^j(x,y) * A_L^j(x,y) + (1 - d^j(x,y)) * B_L^j(x,y)$, where $F_L^j(x,y)$ represents the low-frequency sub-image after the fusion of $A_L^j(x,y)$ and $B_L^j(x,y)$ at scale j. In Equation (14), T is the matching threshold between the pixel A and pixel B of a multi-focus image.

High-Frequency Component Fusion Rules

In Equation (8), the first row of S represents low-frequency component of the original image, which carries the primary information from the image. The other rows $S(2:mn,:)$ of S denotes the high-frequency components of the original image, presenting the details of the image. According to the orthogonality of singular value decomposition, each component forms an orthogonal complement on the same scale. The direct sum of each component is:

$$I_j = I_{j+1} \oplus \sum\nolimits_{i=2}^{mn} S(i,:)_{j+1}\ (j = 2,1,0) \tag{15}$$

where j represents the decomposition scale; when $j=2$, the highest decomposition layer is 3, $I_3 = S(1,:)_3$, and each component can be written as:

$$\begin{cases} I_2 = S(1,:)_3 \oplus \sum_{i=2}^{mn} S(i,:)_3 j = 2, \\ I_1 = I_2 \oplus \sum_{i=2}^{mn} S(i,:)_2 j = 1, \\ I_0 = I_1 \oplus \sum_{i=2}^{mn} S(i,:)_1 j = 0, \end{cases} \tag{16}$$

The high-frequency sub-images of QMSVD reflect the detailed characteristics of the original image. Most of the fused methods operate in the feature domain of high-frequency components, without taking the influence of low frequency into account, compromising the fusion quality. To factor in the influence of low-frequency components in high-frequency component fusion, a local contrast fusion rule, which is applicable to both high-frequency and low-frequency regions, is proposed. After the original image is decomposed by QMSVD, the local contrast of the high-frequency and low-frequency components can be obtained by the following equation (Pu and Ni, 2000):

$$C_{S_j}^k(x,y) = I_{S_j}^{H_k}(x,y)/I_{AB_j}^L(x,y), (S_j = A_j or B_j) \tag{17}$$

In Equation (17), $I_{AB_j}^L$ represents the fusion component of the low-frequency sub-image of the original image A and B at scale j, and $I_{S_j}^{H_k}$ represents the *k-th* high-frequency component of the original image S at scale j. According to Equation (15), the high-frequency is not aliased with low-frequency components, and therefore the definition of the local contrast mirroring the high-frequency components is valid. The high-frequency sub-image fusion is defined as:

$$H_{F_j}^k(x,y) = \begin{cases} I_{A_j}^{H_k}(x,y), if\ \left|C_{A_j}^k(x,y)\right| \geq \left|C_{B_j}^k(x,y)\right| \\ I_{B_j}^{H_k}(x,y), otherwise \end{cases} \tag{18}$$

where $H_{F_j}^k(x,y)$ represents the kth high-frequency component of the fused image F at scale j.

Multi-Focus Color Image Fusion Process

Figure 2 shows the scheme of multi-focus color image fusion based on QMSVD with six channels, and the corresponding fusion process is as follows:

Step 1: Two original color multi-focus images A and B are decomposed by QMSVD. The low-frequency sub-image A_L, B_L is represented by one channel and the high-frequency sub-images A_{Hi}, B_{Hi} (H_i is the ith high-frequency channel) are represented by multiple channels. The orthogonal matrices U_A and U_B, corresponding to singular values, are also obtained.

Step 2: The low-frequency sub-images A_L, B_L are fused following low-frequency fusion rules, and the high-frequency sub-images A_{Hi}, B_{Hi} are fused using high-frequency fusion rules.

Step 3: The orthogonal matrices U_A and U_B (obtained in Step 1) are fused. In the fusion of two images after QMSVD decomposition, the roles of U_A and U_B are identical, so the fusion rule for the orthogonal matrix is: $U_F = (U_A + U_B)/2$.

Step 4: The final fusion image is obtained by inverse QMSVD transform of the fusion results in Step 2 and Step 3.

EXPERIMENTAL RESULTS AND DISCUSSION

In this study, color information richness (CCM) (Yuan et al., 2011), spatial frequency (SF), image contrast metric (ICM) (Yuan et al., 2011), and edge information retention (QAB/F) (Liu et al., 2012) are utilized to evaluate the multi-focus color fusion image objectively, and to verify the effectiveness of the algorithm. The CCM index value is determined by the color chromaticity and color difference gradient of the fused image. The SF index reflects the clarity of the image details. The ICM index is composed of the grayscale contrast and color contrast of the fused image, with the value denoting the contrast in the fused image. The QAB/F index implies how much information about edge and structure from the original image is retained in the fused image. For the above evaluation indicators, a larger evaluation value suggests a better fusion result.

The proposed QMSVD color image fusion method is compared with five typical multi-focus image fusion methods, which fall into the category of the multi-resolution singular value decomposition fusion method (MSVD) (Naidu, 2011), the Multi-scale weighted gradient-based fusion method (MWGF) (Zhou et al., 2014), the boosted random walks-based fusion method (RWTS) (Ma et al., 2019), the guided fifilter-based fusion method (GFDF) (Qiu et al., 2019), the deep CNN fusion method (CNN) (Liu et al., 2017). Among them, the MSVD, MWGF, RWTS and GFDF are traditional image fusion methods. The CNN is a recently proposed image fusion method based on deep learning. In Liu et al. (2017), Liu chooses the Siamese as the CNN model, and the network has three convolutional layers and one max-pooling layer. The training sample is a high-quality natural image of 50,000 from the ImageNet dataset, and input patch size is set at 16 × 16. The Matlab implementation of the above five fusion methods are all obtained online, and the parameters are the default values given in the literature. The original multi-focus images used in the experiment are obtained from multiple image datasets. The four images (A), (B), (D), (E) in **Figure 4** and the one image (I) in **Figure 6** are obtained from the Lytro dataset (Nejati et al., 2015). The Six images (A)–(F) in **Figure 6** are obtained from the Slavica dataset (Slavica, 2011). The one image (C) in **Figure 4** and the two images (G) and (H) in **Figure 6** are obtained from the Saeedi dataset (Saeedi and Faez, 2015). The one image (J) in **Figure 6** is obtained from the Bavirisetti dataset (Bavirisetti). In this paper, five groups of color images with rich colors are selected in the image datasets Lytro and Saeedi, and they are used in the comparison experiment. In addition, 10 groups multi-focus images commonly used in other related papers as the experimental data are used in the comparison experiment, and they have different sizes and characteristics.

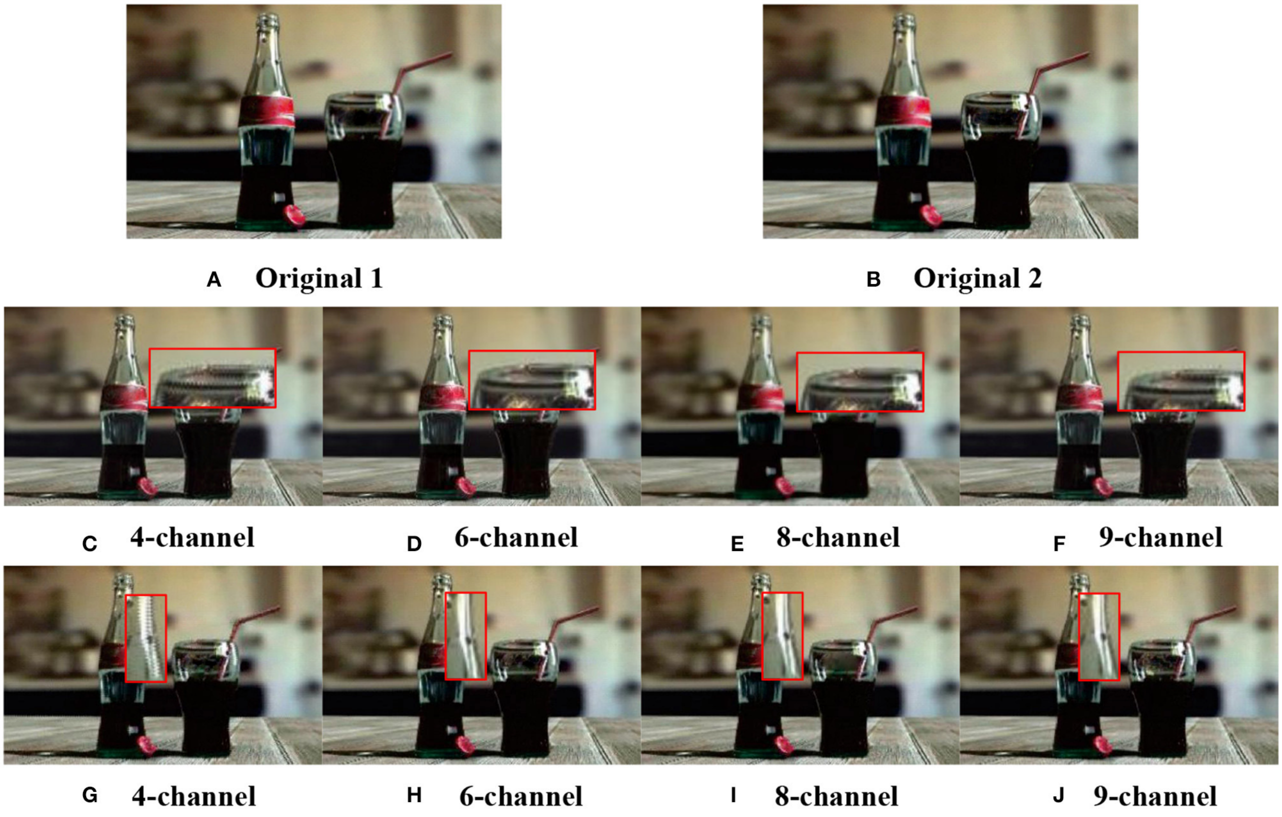
In the experimental process, firstly, the experimental parameters of the algorithm set prior to the experiment. Secondly, the fusion results achieved using the proposed algorithm and the other algorithms are presented and compared.

Selection of Experimental Parameters

Multi-scale singular value decomposition of color images is conducted through multiple independent layers and channels. Image decomposition generally divides the image into three layers. Channel decomposition usually divides the image into four-channel, six-channel, eight-channel, and nine-channel. Channel decomposition is illuminated in Equation (9). The result of image fusion is also affected related to the size of the local window P, and the typical size is 3×3 or 5×5. The experimental comparison suggests, the 5×5 local window exceeds the size of the important feature of the image, which undermines the judgment of the local window activity. Therefore, in this paper, we set a local window size at $P = 3 \times 3$. As can be observed from Equation (11), the weight $t \in [0, 1]$ indicates the relative importance of chromaticity and luminance, with t positively related to chromaticity. In Equation (14), T represents the matching threshold of the matching level between the pixels of the two color multi-focus images to be fused, and the value of T directly affects the decision value $d(x, y)$ of low-frequency fusion. The parameters discussed above ultimately determine the effect of image fusion.

We set different parameter values, conducted repeated comparative experiments, and used two objective indices spatial frequency (SF) and color colorfulness metric (CCM) (Yuan et al., 2011) to evaluate **Figure 3**. As **Table 1** reveals, the SF value decreases as the number of channels increases, the larger the number of channels the smoother the image after multi-scale singular value decomposition, and the lower the spatial frequency. The maximum value of CCM occurs when $t = 0.9$. According to Equation (11), value t indicates the importance of chromaticity. The analysis shows that the algorithm proposed in this paper is feasible. From further analysis in **Table 1**, the preliminary parameters could be obtained: $P = 3$, $t = 0.9$, $T = 0.01$, and $P = 3$, $t = 0.9$, $T = 0.03$, with six and eight decomposition channels.

Figure 3 demonstrates the results obtained in the second decomposition layer using the preliminary parameters analyzed above. Obviously, the fusion image based on four channel



A Original 1

B Original 2

C 4-channel

D 6-channel

E 8-channel

F 9-channel

G 4-channel

H 6-channel

I 8-channel

J 9-channel

FIGURE 3 | Tested multi-focus color image. **(A,B)** are original images. The parameters are selected: in **(C–F)**, layer = 2, $P = 3$, $t = 0.9$, $T = 0.01$; in **(G–J)**, layer = 2, $P = 3$, $t = 0.9$, $T = 0.03$; with four, six, eight, and nine decomposition channels.

TABLE 1 | Selection of initial parameters (1).

Parameters	Decompose	Spatial frequency (SF)				Color colorfulness metric (CCM)			
		4-channel	6-channel	8-channel	9-channel	4-channel	6-channel	8-channel	9-channel
$t = 0.8$, $T = 0.01$	1-layer	27.7544	27.1201	27.0089	24.0205	17.1906	17.2659	17.3272	17.1945
	2-layer	27.7476	27.1008	26.9451	23.9930	17.2387	17.3145	17.3351	17.1784
	3-layer	27.7046	27.0782	26.9962	24.0224	17.2168	17.3000	17.2959	17.1949
$t = 0.8$, $T = 0.02$	1-layer	27.7530	27.1597	27.0169	24.0233	17.1766	17.2540	17.3045	17.1713
	2-layer	27.7565	27.0789	26.9332	23.9857	17.1433	17.2997	17.3361	17.1717
	3-layer	27.7017	27.0856	26.9821	23.9763	17.1764	17.2824	17.2891	17.1752
$t = 0.8$, $T = 0.03$	1-layer	27.7701	27.1563	27.0112	24.0128	17.1863	17.2646	17.3160	17.1813
	2-layer	27.7474	27.0726	26.9467	24.0115	17.1571	17.2842	17.3379	17.1754
	3-layer	27.6838	27.0666	26.9722	23.9911	17.1753	17.2689	17.2997	17.1662
$t = 0.9$, $T = 0.01$	1-layer	27.7659	27.1194	**27.0250**	24.0425	17.1872	17.2721	17.3282	17.2052
	2-layer	27.7655	27.0759	26.9507	24.0050	17.2252	**17.3203**	17.3368	17.1696
	3-layer	27.7285	27.0611	26.9984	24.0227	17.2239	17.3126	**17.3529**	17.1971
$t = 0.9$, $T = 0.02$	1-layer	27.7560	27.1401	27.0168	24.0097	17.1729	17.2506	17.2962	17.1755
	2-layer	27.7343	27.0708	26.9335	23.9877	17.1538	17.3002	17.3354	17.1735
	3-layer	27.6954	27.0692	26.9641	23.9779	17.1793	17.2707	17.2944	17.1755
$t = 0.9$, $T = 0.03$	1-layer	27.7818	**27.1624**	27.0196	23.9995	17.1861	17.2599	17.3134	17.1780
	2-layer	27.7563	27.0767	26.9629	24.0269	17.1555	17.2871	17.3332	17.1750
	3-layer	27.7052	27.0612	26.9701	24.0058	17.1757	17.2689	17.3034	17.1634

The numbers in bold indicate the maximum value obtained with different objective evaluation indicators.

TABLE 2 | Selection of initial parameters (2).

Channel	Metrics	$t = 0.8$, $T = 0.01$	$t = 0.8$, $T = 0.02$	$t = 0.8$, $T = 0.03$	$t = 0.9$, $T = 0.01$	$t = 0.9$, $T = 0.02$	$t = 0.9$, $T = 0.03$	Total
6-channel	SF	81.2991	81.3242	81.2955	81.2564	81.2801	81.3003	798.8322
	CCM	51.8804	51.8361	51.8177	51.905	51.8215	51.8159	
8-channel	SF	51.9582	51.9297	51.9536	52.0179	51.926	51.95	797.389
	CCM	80.9502	80.9322	80.9301	80.9741	80.9144	80.9526	

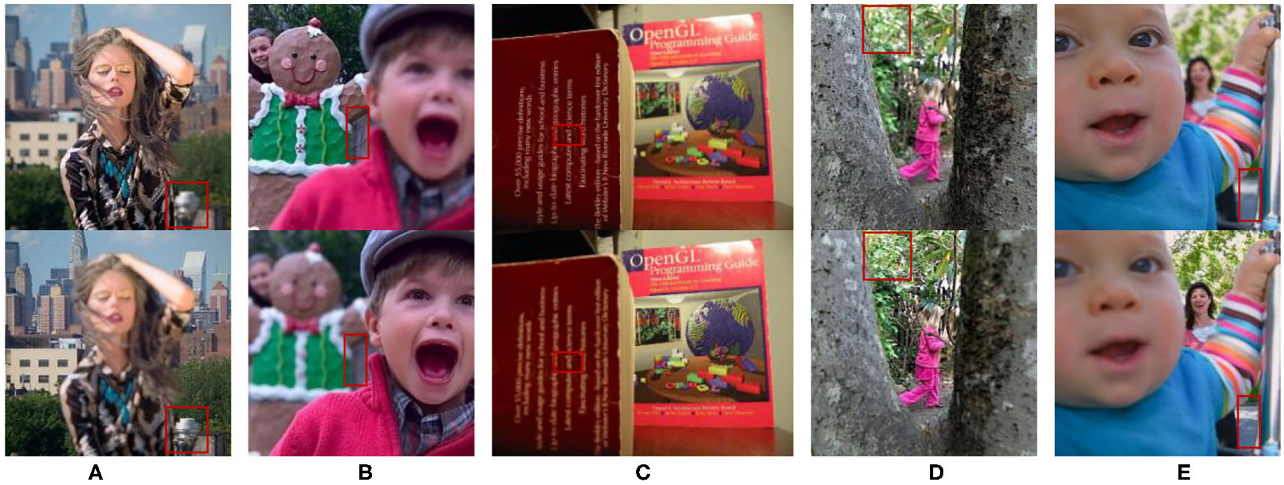


FIGURE 4 | Five groups of multi-focus color original images. Red frames are the area that need to be compared in image fusion. **(A)** Woman, **(B)** Child, **(C)** Book, **(D)** Girl, and **(E)** Baby. The four images **(A,B,D,E)** from Lytro dataset, the image **(C)** from Saeedi dataset.

decomposition has the worst visual effect, and the edge of detail appears zigzag distortion, which results from the block effect caused by small channel decomposition. Artifacts emerge at the edge of fused image obtained through nine- channel decomposition. This due to the large channel decomposition which lead to blurring of the fused image. Fused images obtained through six-channel and eight-channel decomposition have similar effects and the best quality. Judging from the **Table 1**, it can be concluded that the subjective visual effects are consistent with the objective evaluation values. In other words, the objective evaluation value is positively proportional to the subjective visual effect.

From the analysis above, the fusion effects of the six-channel and eight-channel decomposition are superior to those of the four-channel or nine-channel decomposition. Further analysis from **Table 2** reveals that the overall results of SF and CCM with six channels are better than those with eight channels, therefore, we finally adopt the six-channel decomposition approach. According to **Table 1**, during the six-channel decomposition, when $P = 3$, $t = 0.9$, $T = 0.03$, and layer = 1, the maximum SF value is 27.1624, and when layer = 2, the maximum CCM value is 17.2871. To optimize the result of multi-focus color image fusion, we take into account importance of color evaluation index CCM in color image fusion, and take the six-channel decomposition approach, and set $P = 3$, $t = 0.9$, $T = 0.03$, and layer = 2.

Subjective Evaluation

To verify the performance of the proposed method of multi-focus color image fusion in terms of visual perception, 15 groups of multi-focus color images are selected for our experiment. Five groups come from the multi-focus image data set "Lytro," while the other 10 groups are widely used in multi-focus image fusion. Meanwhile, the proposed fusion method is compared with five typical multi-focus image fusion methods, which are the MSVD, MWGF, RWTS, GFDF and CNN.

In **Figure 4**, we select five groups images from the multi-focus data set "lytro" for experiments. They have rich colors, which are also the experimental data used in the five comparison algorithms. The areas in each image that need to be compared are marked with a red frame. **Figure 5** is the fusion result corresponding to the five original images in **Figure 4**. For better comparison, the red frame areas in the fusion image are enlarged.

Group A(1)–A(6) show the images of the "woman" with the size of 208 × 208 and the fused image obtained by 6 different fusion methods. The comparison of red framed areas suggest the QMSVD, RWTS, MSVD, and GFDF have the best visual clarity, followed by CNN, and MDGF is the most blurry. A further comparison shows that in the fused image obtained by the MSVD, the red framed region and the image of "woman" have obvious color distortion.

Group B(1)–B(6) show the images of the "child" with the size of 256 × 256 and the fused image obtained by six different fusion methods. The comparison of the red framed areas demonstrates that the QMSVD and MWGF have the best visual clarity, and GFDF is the fuzziest. A further comparison shows that in the fusion image obtained by the MSVD, the face brightness of "child" is the lowest.

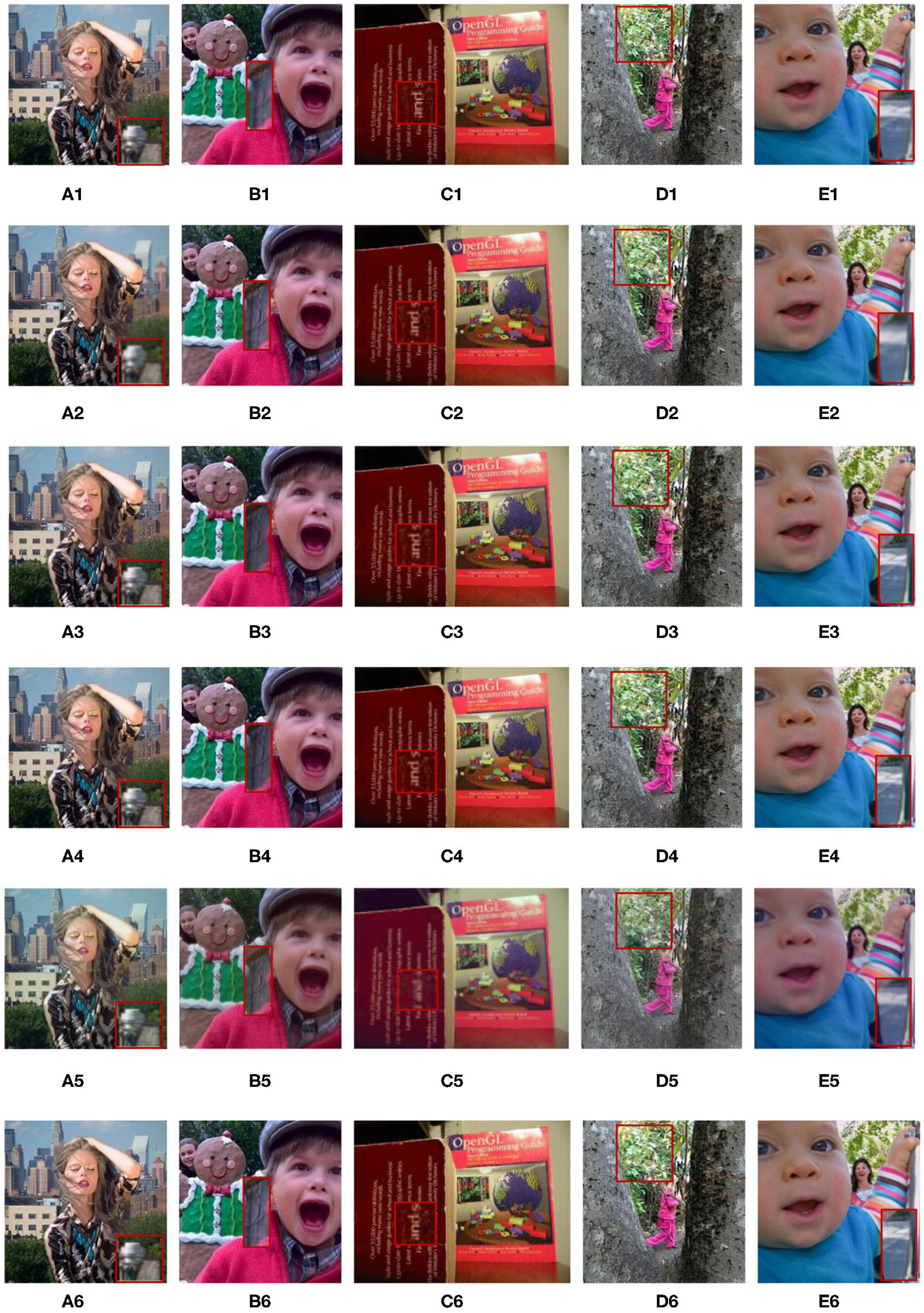


FIGURE 5 | Corresponding to the fusion results of the five original images in **Figure 4**. **A(1)–E(1)** are the fusion images obtained by the GFDF method. **A(2)–E(2)** are the fusion images obtained by the MWGF method. **A(3)–E(3)** are the fusion images obtained by the CNN method. **A(4)–E(4)** are the fusion images obtained by the RWTS method. **A(5)–E(5)** are the fusion images obtained by the MSVD method. **A(6)–E(6)** are the fusion images obtained by the QMSVD method.

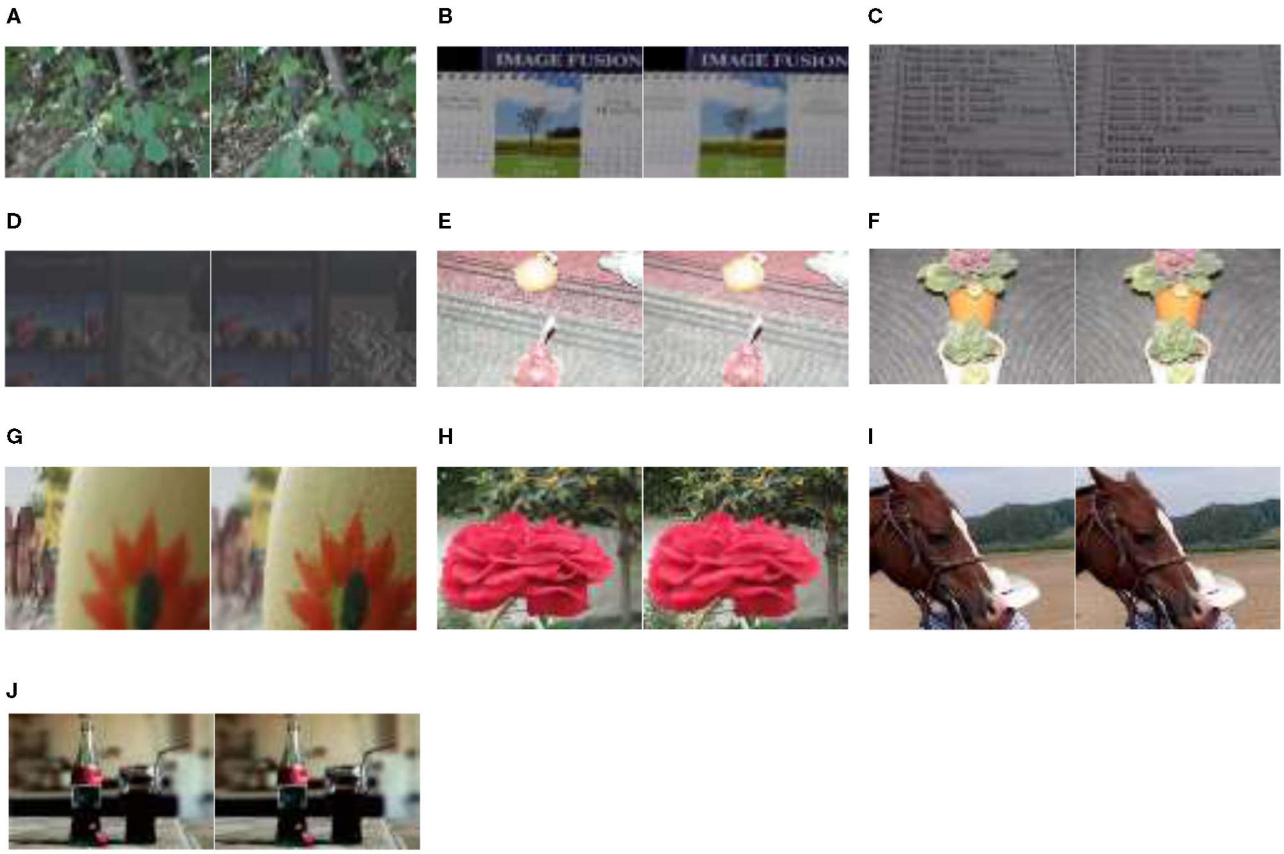


FIGURE 6 | Ten groups of multi-focus color images. **(A)** Size of 267×171, **(B)** size of 267×175, **(C)** size of 267×177, **(D)** size of 267×177, **(E)** size of 267×174, **(F)** size of 320×200, **(G)** size of 267×174, **(H)** size of 390×260, **(I)** size of 222×148, and **(J)** size of 360×360. The six images **(A–F)** from Slavica dataset, the two images **(G,H)** from Saeedi dataset, the image **(I)** from Lytro dataset and the image **(J)** from Bavirisetti dataset.

Group C(1)–C(6) show the images of the "book" with the size of 320 × 240 and the fused image obtained by six different fusion methods. Comparing the English letters in the red frame area of each image. From a visual point of view, the MSVD-based method is the most blurry, and fusion effects achieved by the other methods are similar.

Group D(1)–D(6) show the images of the "girl" with the size of 300 × 300 and the fused image obtained by six different fusion methods. Comparing the leaves in the red frame area of each image, the QMSVD and the RWTS can produce the best fusion image effect, and the color is close to the original image.

Group E(1)–E(6) show the images of the "baby" with the size of 360 × 360 and the fused image obtained by 6 different fusion methods. The comparison illustrates that the QMSVD, CNN, and RWTS obtain the best fusion image effects, followed by the GFDF and MSVD, and the MWGF lags behind.

To further prove the effectiveness of the QMSVD method for multi-focus color image fusion, the 10 groups of original images are given in **Figure 6**. In **Figure 7**, the fused image obtained by six different fusion methods are shown. In **Figures 8**, **9**, we compare two groups of images in detail.

In **Figure 8**, the original image of "Coke Bottle" with a size of 320 × 200 and the fused image obtained by six different fusion methods are shown. Compare the bright spots in the red frame area of each image, QMSVD, CNN, and GFDF achieve better clarity, followed by MWGF and RWTS, and MSVD is the most ambiguous.

In **Figure 9**, the original image of "Forest" with a size of 267 × 171 and the fused image obtained by six different fusion methods are shown. Compare the brightness of leaves in the red frame area of each image, QMSVD, superior to other methods, obtains the best fusion image effect.

In general, the QMSVD method combines the advantages of quaternions and multi-scale decomposition in color multi-focus image fusion. The benefit is that quaternions can represent and process different color channels of a color image as a whole, producing the fused multi-focus image with high fidelity. Multi-scale decomposition methods decompose the image into low-frequency and high-frequency components at different levels. In this way, the decomposed images can be fused accurately at different components, scales, and levels, which renders the fused color multi-focus image with high definition and contrast, and good visual effects.

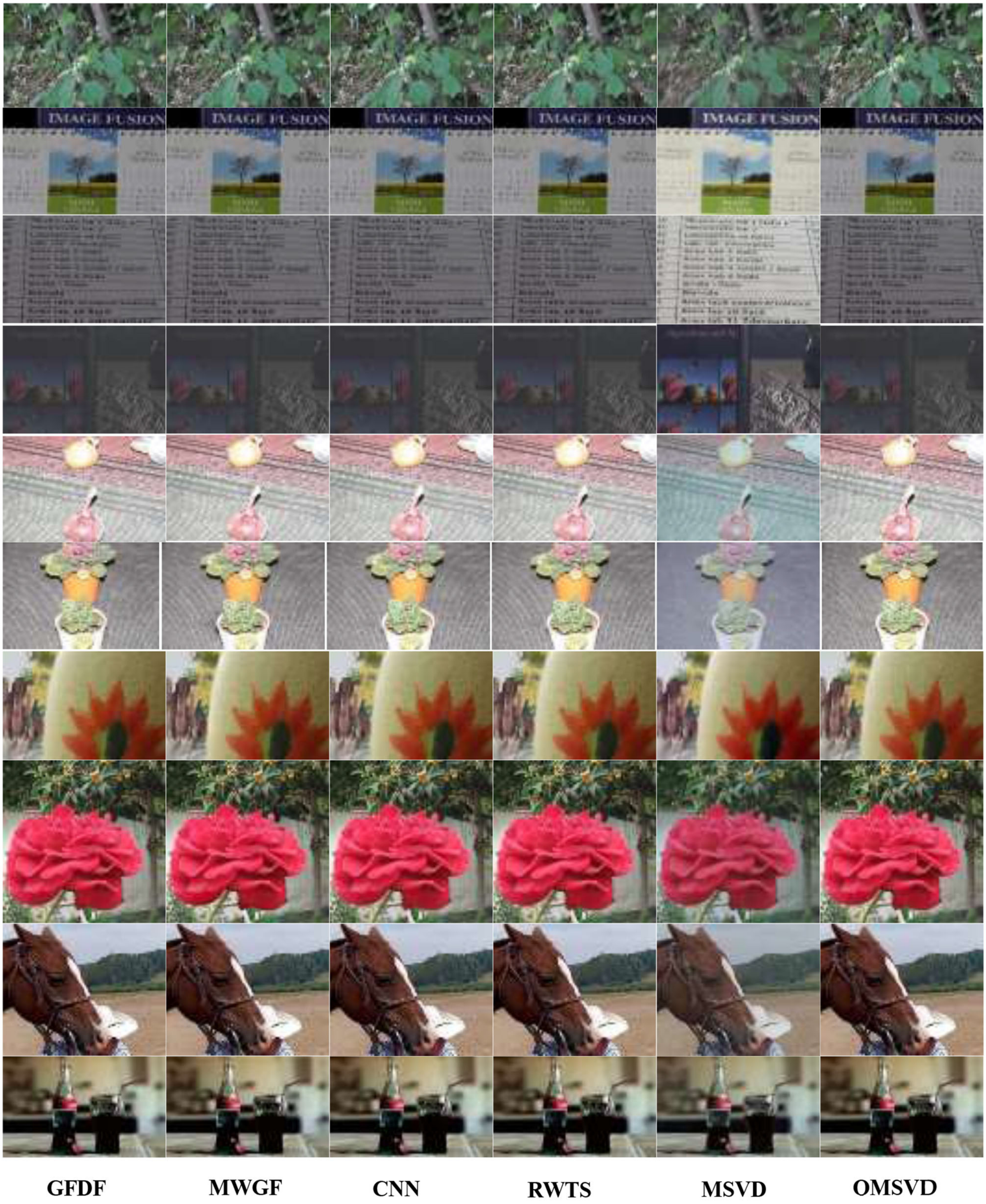


FIGURE 7 | Ten groups of multi-focus color fusion images.

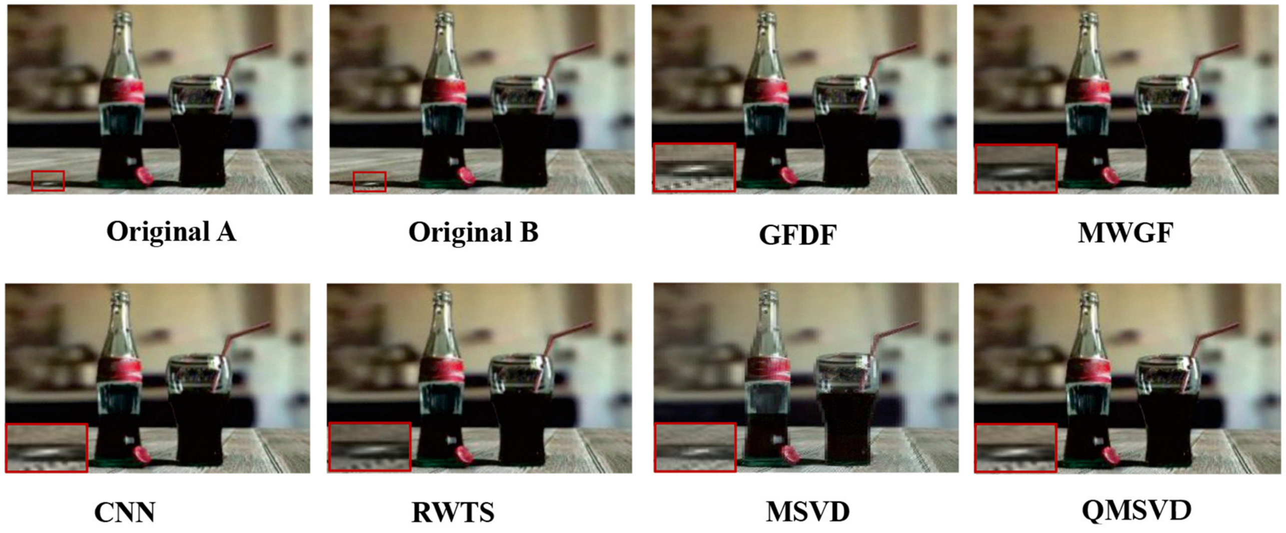


FIGURE 8 | "Coke Bottle" fusion images obtained by six different fusion methods.

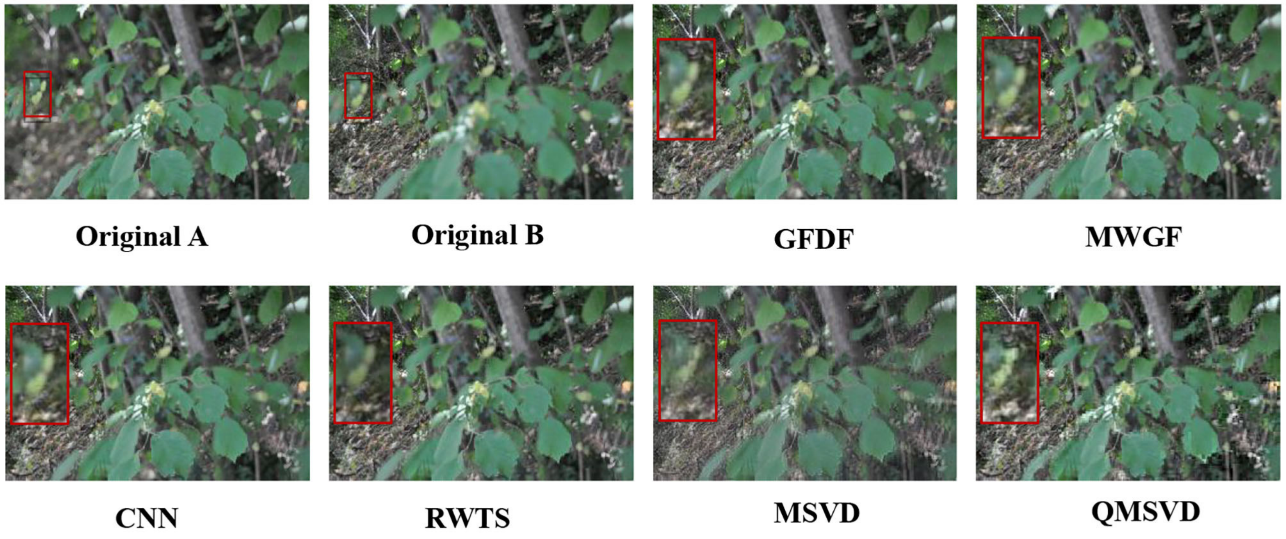


FIGURE 9 | "Forest" fusion images obtained by six different fusion methods.

Objective Evaluation

We proposed the method for multi-focus color image fusion. We classify experimental images in two categories. One type is multi-focus color pictures with rich color information, and their objective evaluation metrics of different methods are presented in **Table 3**. The other type is commonly used multi-focus color images. We have selected two groups, and their objective evaluation metrics of different methods are counted in **Table 4**. **Table 5** is the average objective evaluation metrics of different methods on 15 groups color images. The analysis of **Tables 3–5** shows that the average values of the 15 groups using CCM and ICM indicators of the QMSVD algorithm are significantly higher than those of other fusion algorithms. This also shows that the fused image has a high definition and rich color, which is consistent with the visual performance of the fused image in the subjective evaluation. Of all the fusion algorithms, the CCM index of the QMSVD algorithm ranks first. For the QAB/F indicator, the QMSVD algorithm performs worse than other algorithms in preserving edge and structure information. In general, the QMSVD method achieves the best results on the CCM indicator and performs well on the ICM and SF indicators. This shows that the QMSVD method is effective, and the fused image has a high definition, rich color, less information loss, and good overall visual effects.

TABLE 3 | Objective evaluation values of multi-focus color images.

Image	Metrics	GFDF	MWGF	RWTS	CNN	MSVD	QMSVD	Rank
“Woman”	CCM	19.8116	19.6341	19.8065	19.8219	19.2142	**19.9050**	1
	ICM	0.5448	**0.5555**	0.5447	0.5451	0.5463	0.5538	2
	SF	30.2661	29.4158	30.3256	29.8675	27.3840	**30.8316**	1
	QAB/F	0.6845	0.6656	0.6847	**0.6866**	0.6523	0.6692	4
“Child”	CCM	26.6408	26.5463	26.6100	26.5925	25.5832	**26.7334**	1
	ICM	0.4910	0.4913	0.4912	0.4915	0.3638	**0.4988**	1
	SF	25.1688	24.9778	24.8987	24.6005	18.9458	**25.4489**	1
	QAB/F	0.6240	0.6202	**0.6251**	0.6248	0.5054	0.5955	5
“Book”	CCM	28.7924	28.7851	28.7922	28.7846	26.9576	**28.9861**	1
	ICM	0.4582	0.4578	0.4578	0.4578	0.3506	**0.4610**	1
	SF	35.3490	**35.5172**	35.3293	35.2197	17.5239	33.9891	5
	QAB/F	0.6832	0.6814	0.6848	**0.6853**	0.3768	0.5944	5
“Girl”	CCM	20.5994	20.5039	20.5796	20.5546	17.5183	**20.6437**	1
	ICM	0.5311	0.5313	0.5312	0.5317	0.4446	**0.5340**	1
	SF	48.7194	48.5660	48.3491	47.8869	35.3703	**48.8196**	1
	QAB/F	0.6992	0.6943	0.7015	**0.7023**	0.6260	0.6854	5
“Baby”	CCM	24.9107	24.8468	24.9080	24.9040	16.6895	**24.9657**	1
	ICM	0.5161	0.5377	0.5161	0.5162	0.4229	**0.5739**	1
	SF	**19.4409**	19.1723	19.3729	19.2464	13.3334	19.3610	3
	QAB/F	0.6682	0.6599	0.6701	**0.6712**	0.5066	0.6479	5

The numbers in bold indicate the maximum value obtained with different objective evaluation indicators.

TABLE 4 | Objective evaluation metrics of multi-focus color images in **Figures 8**, **9**.

Image	Metrics	GFDF	MWGF	RWTS	CNN	MSVD	QMSVD	Rank
“Coke Bottle”	CCM	17.2691	17.2181	17.2865	17.2782	**15.1438**	**17.2871**	1
	ICM	0.5521	**0.5523**	0.5521	0.5521	0.4400	0.5508	2
	SF	**27.5118**	27.0422	27.4867	27.4254	19.1469	27.0767	4
	QAB/F	0.7609	0.7446	0.7609	**0.7613**	0.4820	0.7563	4
“Forest”	CCM	21.2723	21.2740	21.2442	21.2616	20.7242	**21.5211**	1
	ICM	0.4493	0.4496	0.4503	0.4495	0.4346	**0.5120**	1
	SF	26.5008	26.4351	26.6436	26.3499	23.4413	**29.6777**	1
	QAB/F	**0.6232**	0.6229	0.6188	0.6171	0.4182	0.4626	5

The numbers in bold indicate the maximum value obtained with different objective evaluation indicators.

TABLE 5 | Average objective evaluation metrics of different methods on 15 groups color images.

Image	Metrics	GFDF	MWGF	RWTS	CNN	MSVD	QMSVD	Rank
15 groups color images	**CCM**	20.0358	19.9940	20.0308	20.0249	20.5252	**21.4558**	1
	ICM	0.4606	0.4641	0.4599	0.4581	0.3558	**0.4763**	1
	SF	**28.4214**	28.2365	28.3956	28.1854	23.6767	28.3095	3
	QAB/F	**0.6821**	0.6713	0.6820	0.6818	0.4619	0.6030	5

The numbers in bold indicate the maximum value obtained with different objective evaluation indicators.

CONCLUSION

In this paper, a multi-focus color image fusion algorithm based on quaternion multi-scale singular value decomposition is proposed. In the algorithm, the color multi-focus image, represented by quaternions, undergoes multi-scale decomposition as a whole, avoiding the loss of color information caused by the multi-scale decomposition of each color channel separately. In addition, the algorithm can fuse the information of the decomposed image accurately in different components, scales, and levels. To verify the effectiveness of the algorithm, it has been analyzed qualitatively and quantitatively, and compared with the classical multi-scale decomposition fusion algorithm and fusion algorithms proposed in the latest literature. The experimental results show that the fusion result of this method reports great enhancement in the subjective visual effects. It also performs well in objective evaluation indices, particularly the CCM index of color information richness of the fused image. Because the algorithm proposed in this paper is based on multi-focus color images represented by quaternion, it takes more time to process the multi-scale decomposition of the images. Further research needs to be done to improve the efficiency of the algorithm and ensure the quality of image fusion. Regarding the setting of algorithm parameters, it is mainly based on empirical values, such as the selection of the number of channels, the selection of local window size, etc. In the future, the adaptive selection of parameters is also the focus of our future research. Additionally, the color images are not represented by the complete quaternion components, but by pure quaternion in image fusion. How to exploit the real part information of quaternion in color image processing will be our focus in the future study.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://dsp.etfbl.net/mif/; https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset.

AUTHOR CONTRIBUTIONS

HW, XT, and BX conceived this research. HW and BX designed the algorithm. HW performed the computer simulations and wrote the original draft. HW and ZZ analyzed the data. WL and XT revised and edited the manuscript. All authors confirmed the submitted version.

FUNDING

This research was funded by the National Nature Science Foundation of China under Project 61673079, the Natural Science Foundation of Chongqing under Project cstc2018jcyjAX0160, and the Educational Commission Foundation of Chongqing of China (Grant Nos. KJQN201900547 and KJ120611).

ACKNOWLEDGMENTS

The authors thank the editors and the reviewers for their careful works and valuable suggestions for this study.

REFERENCES

An, F., and Li, Z. (2019). Image processing algorithm based on bi-dimensional local mean decomposition. *J. Math. Imaging VIS* 61, 1243–1257. doi: 10.1007/s10851-019-00899-8

Aymaz, S., and Köse, C. (2019). A novel image decomposition-based hybrid technique with super-resolution method for multi-focus image fusion. *Inf. Fusion* 45, 113–127. doi: 10.1016/j.inffus.2018.01.015

Bavirisetti, D. P. *Fusion Image Dataset*. Available online at: https://sites.google.com/view/durgaprasadbavirisetti/datasets

Bihan, N., and Sangwine, S. (2003). "Color image decomposition using quaternion singular value decomposition," in *Proc. 2003 Int. Conf.Visual Information Engineering (VIE 2003)* (Surrey), 113–116.

Burt, P., and Kolczynski, R. (1993). "Enhanced image capture through fusion," in *Fourth IEEE International Conference on Computer Vision* (Berlin), 173–182.

Chen, B., Shu, H., Coatrieux, G., Chen, G., and Coatrieux, J. (2014). Color image analysis by quaternion-type moments. *J. Math. Imaging VIS* 51, 124–144. doi: 10.1007/s10851-014-0511-6

Du, J., Li, W., Xiao, B., and Nawaz, Q. (2016). Union Laplacian pyramid with multiple features for medical image fusion. *Neurocomputing* 194, 326–339. doi: 10.1016/j.neucom.2016.02.047

Duan, J., Chen, L., and Chen, C. (2018). Multifocus image fusion with enhanced linear spectral clustering and fast depth map estimation. *Neurocomputing* 318, 43–54. doi: 10.1016/j.neucom.2018.08.024

Gonzalo, P., and Jesús, M. (2002). A wavelet-based image fusion tutorial. *Pattern Recognit.* 37, 1855–1872. doi: 10.1016/j.patcog.2004.03.010

Grigoryan, A. M., and Agaian, S. S. (2018). *Quaternion and Octonion Color Image Processing With Matlab*. Bellingham, WA: SPIE, 111–139.

Jaroslav, K., Jan, F., Barbara, Z., and Stanislava, S. (2002). A new wavelet-based measure of image focus. *Pattern Recognit. Lett.* 23, 1785–1794. doi: 10.1016/S0167-8655(02)00152-6

Jin, L., Song, E., Li, L., and Li, X. (2013). "A quaternion gradient operator for color image edge detection," in *IEEE International Conference on Image Processing (ICIP)* (Melbourne, VIC), 3040–3044.

Jin, X., Chen, G., Hou, J., Jiang, Q., Zhou, D., and Yao, S. (2018a). Multimodal sensor medical image fusion based on nonsubsampled shearlet transform and S-PCNNs in HSV space. *Signal Process.* 153, 379–395. doi: 10.1016/j.sigpro.2018.08.002

Jin, X., Jiang, Q., Chu, X., Xun, L., Yao, S., Li, K., et al. (2020). Brain medical image fusion using L2-norm-based features and fuzzy-weighted measurements in 2D littlewood-paley EWT domain. *IEEE Trans. Instrum. Meas.* 69, 5900–5913. doi: 10.1109/TIM.2019.2962849

Jin, X., Zhou, D., Yao, S., and Nie, R. (2018b). Multi-focus image fusion method using S-PCNN optimized by particle swarm optimization. *Soft Comput.* 22, 6395–6407. doi: 10.1007/s00500-017-2694-4

Kakarla, R., and Ogunbona, P. (2001). Signal analysis using a multiresolution form of the singular value decomposition. *IEEE Trans. Image Process.* 10, 724–735. doi: 10.1109/83.918566

Li, J., Guo, X., Lu, G., Zhang, B., Xu, Y., Wu, F., et al. (2020). Drpl: deep regression pair learning for multi-focus image fusion. *IEEE Trans. Image Process.* 29, 4816–4831. doi: 10.1109/TIP.2020.2976190

Li, S., Kang, X., Fang, L., Hu, J., and Yin, H. (2017). Pixel-level image fusion: a survey of the state of the art. *Inf. Fusion* 33, 100–112. doi: 10.1016/j.inffus.2016.05.004

Li, Y., Sun, Y., Huang, X., Qi, G., Zheng, M., and Zhu, Z. (2018). An image fusion method based on sparse representation and sum modified-laplacian in NSCT domain. *Entropy* 20:522. doi: 10.3390/e20070522

Liang, J., and He, Y. (2012). Image fusion using higher order singular value decomposition. *IEEE Trans. Image Process.* 21, 2898–2909. doi: 10.1109/TIP.2012.2183140

Liu, S., Hu, Q., Li, P., Zhao, J., Liu, M., and Zhu, Z. (2019a). Speckle suppression based on weighted nuclear norm minimization and grey theory. *IEEE Trans. Geosci. Remote Sens.* 57, 2700–2708. doi: 10.1109/TGRS.2018.2876339

Liu, S., Ma, J., Yin, L., and Hu, S. (2020). Multi-focus color image fusion algorithm based on super-resolution reconstruction and focused area detection. *IEEE Access* 8, 90760–90778. doi: 10.1109/ACCESS.2020.2993404

Liu, S., Wang, J., Lu, Y., Hu, S., Ma, X., and Wu, Y. (2019b). Multi-focus image fusion based on adaptive dual-channel spiking cortical model in non-subsampled Shearlet domain. *IEEE Access* 7, 56367–56388. doi: 10.1109/ACCESS.2019.2900376

Liu, Y., Chen, X., Peng, H., and Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Information Fusion.* 36, 191–207. doi: 10.1016/j.inffus.2016.12.001

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020). Multi-focus image fusion: A Survey of the state of the art. *Information Fusion.* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Liu, Z., Blasch, E., Xue, Z., Zhao, J., Laganiere, R., and Wu, W. (2012). Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 94–109. doi: 10.1109/TPAMI.2011.109

Ma, J., Zhou, Z., Wang, B., Miao, L., and Zong, H. (2019). Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps. *Neurocomputing* 335, 9–20. doi: 10.1016/j.neucom.2019.01.048

Naidu, V. P. S. (2011). Image fusion technique using multi-resolution singular value decomposition. *Def. Sci. J.* 61, 479–484. doi: 10.14429/dsj.61.705

Nejati, M., Samavi, S., and Shirani, S. (2015). *Multi-Focus Image Dataset*. Available online at: https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset

Pu, T., and Ni, G. (2000). Contrast-based image fusion using the discrete wavelet transform. *Opt. Eng.* 39, 2075–2082. doi: 10.1117/1.1303728

Qiu, X., Li, M., Zhang, L., and Yuan, X. (2019). Guided filter-based multi-focus image fusion through focus region detection. *Signal Process Image* 72, 35–46. doi: 10.1016/j.image.2018.12.004

Saeedi, J., and Faez, K. (2015). *Multi-Focus Image Dataset*. Technical Report. Available online at: https://www.researchgate.net/publication/273000238_multi-focus_image_dataset

Slavica, S. (2011). *Multi-Focus Image Dataset*. Available online at: https://dsp.etfbl.net/mif/

Vishal, C., and Vinay, K. (2018). Block-based image fusion using multi-scale analysis to enhance depth of field and dynamic range. *Signal Image Video Process* 12, 271–279. doi: 10.1007/s11760-017-1155-y

Xiao, B., Ou, G., Tang, H., Bi, X., and Li, W. (2020). Multi-focus image fusion by hessian matrix based decomposition. *IEEE Trans. Multimedia* 22, 285–297. doi: 10.1109/TMM.2019.2928516

Xu, Y., Yu, L., Xu, H., Zhang, H., and Nguyen, T. (2015). Vector sparse representation of color image using quaternion matrix analysis. *IEEE Trans. Image Process.* 4, 1315–1329. doi: 10.1109/TIP.2015.2397314

Yuan, Y., Zhang, J., Chang, B., and Han, Y. (2011). Objective quality evaluation of visible and infrared color fusion image. *Opt. Eng.* 50, 1–11. doi: 10.1117/1.3549928

Zhang, Y., Xiang, Z., and Wang, B. (2017). Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure. *Inf. Fusion* 35, 81–101. doi: 10.1016/j.inffus.2016.09.006

Zhou, Z., Li, S., and Wang, B. (2014). Multi-scale weighted gradient-based fusion for multi-focus images. *Inf. Fusion* 20, 60–72. doi: 10.1016/j.inffus.2013.11.005

Zhu, Z., Wei, H., Hu, G., Li, Y., Qi, G., and Mazur, N. (2020). A novel fast single image dehazing algorithm based on artificial multiexposure image fusion. *IEEE Trans. Instrum. Meas.* 70, 1–23. doi: 10.1109/TIM.2020.3024335

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 07 July 2021
doi: 10.3389/fnbot.2021.700483

Single Image Defogging Method Based on Image Patch Decomposition and Multi-Exposure Image Fusion

Qiuzhuo Liu[1,2,3], Yaqin Luo[4], Ke Li[2,3]*, Wenfeng Li[2,3], Yi Chai[1], Hao Ding[2,3] and Xinghong Jiang[2,3]

[1] College of Automation, Chongqing University, Chongqing, China, [2] National Engineering Laboratory for Highway Tunnel Construction Technology, Chongqing, China, [3] China Merchants Chongqing Communications Technology Research & Design Institute Co., Ltd., Chongqing, China, [4] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China

Bad weather conditions (such as fog, haze) seriously affect the visual quality of images. According to the scene depth information, physical model-based methods are used to improve image visibility for further image restoration. However, the unstable acquisition of the scene depth information seriously affects the defogging performance of physical model-based methods. Additionally, most of image enhancement-based methods focus on the global adjustment of image contrast and saturation, and lack the local details for image restoration. So, this paper proposes a single image defogging method based on image patch decomposition and multi-exposure fusion. First, a single foggy image is processed by gamma correction to obtain a set of underexposed images. Then the saturation of the obtained underexposed and original images is enhanced. Next, each image in the multi-exposure image set (including the set of underexposed images and the original image) is decomposed into the base and detail layers by a guided filter. The base layers are first decomposed into image patches, and then the fusion weight maps of the image patches are constructed. For detail layers, the exposure features are first extracted from the luminance components of images, and then the extracted exposure features are evaluated by constructing gaussian functions. Finally, both base and detail layers are combined to obtain the defogged image. The proposed method is compared with the state-of-the-art methods. The comparative experimental results confirm the effectiveness of the proposed method and its superiority over the state-of-the-art methods.

Keywords: image defogging, gamma correction, multi-exposure image fusion, image patch, base and detail layers

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Yafei Zhang,
Kunming University of Science and Technology, China
Kshitij Shinghal,
Moradabad Institute of Technology, India
Jun Li,
Chinese Academy of Sciences (CAS), China

***Correspondence:**
Ke Li
like1@cmhk.com

Received: *26 April 2021*
Accepted: *18 June 2021*
Published: *07 July 2021*

Citation:
Liu Q, Luo Y, Li K, Li W, Chai Y, Ding H and Jiang X (2021) Single Image Defogging Method Based on Image Patch Decomposition and Multi-Exposure Image Fusion. Front. Neurorobot. 15:700483. doi: 10.3389/fnbot.2021.700483

1. INTRODUCTION

In bad weather, small floating particles (such as dust, smoke, etc.) in the air seriously degrade image quality. The color and details of scene are blurred in degraded images (Li Y. et al., 2017), affecting the performance of the applications closely related to image quality, such as outdoor video monitoring, remote sensing, and so on. Therefore, image defogging has become an important application of computer vision.

As a branch of image processing techniques, image defogging techniques can effectively reduce the adverse effects of fog/haze to enhance image contrast and visibility. As shown in **Figures 1A,E** represent two foggy images, and **Figures 1B,F** represent the corresponding fog-free images of **Figures 1A,E**. The heat maps of both foggy and fog-free images are shown in **Figures 1C–H**, respectively. The overall brightness of foggy images **Figures 1C,G** is higher than the corresponding brightness of fog-free images **Figures 1D,H**. Compared with fog-free images, the feature information of foggy images is obviously blurrier, so it is necessary to remove fog/haze for the effective restoration of the captured feature information (Mehrubeoglu et al., 2016). There are many existing image defogging methods, which can be categorized into image enhancement-based, image restoration-based, and image defogging based on deep learning methods.

Most of image restoration-based defogging methods rely on the responses of atmospheric degradation models. These methods need to extract the a priori information of foggy images. Based on the dark channel prior (DCP) method, the a priori law of dark primary color is first obtained by analyzing a large number of haze-free outdoor images, and then the corresponding fog density is estimated (He et al., 2009). Based on single image defogging methods, variable surface shading is added to an atmospheric scattering model. This method assumes that the surface shading and transfer function are statistically independent. According to this assumption, an atmospheric scattering model is analyzed. So, the transfer function is obtained and haze/fog is removed from foggy images (Fattal, 2008). The contrast of input images is enhanced to improve the image visibility (Tan, 2008). In addition, fast image restoration method (Tarel and Hautiere, 2009) and Bayesian defogging method (Nishino et al., 2012; Ju et al., 2019) were proposed. Fog density changes with the depth of scene, so the degradation of image quality also changes in space. Physical degradation models need the corresponding a priori knowledge to obtain the scene depth information. Scene depth information is not only used to estimate the fog/haze distribution, but also affects the defogging performance. The a priori knowledge of physical degradation models can not be directly applied to any scene, so the acquisition of scene depth information is unstable. Without relying on the scene depth information, image enhancement-based defogging methods can effectively achieve image defogging.

With the development of deep learning, deep learning has been applied to image defogging. Image defogging methods based on deep learning are divided into non end-to-end and end-to-end. Non end-to-end methods used convolutional neural network (CNN) to estimate parameters in an atmospheric scattering model and taken parameters as the output. Parameters are introduced into the atmospheric scattering model for image restoration (Cai et al., 2016). End-to-end defogging methods input a foggy image into CNN and the defogged image directly output (Li B. et al., 2017).

Image enhancement-based defogging methods regard image degradation as the lack of contrast and saturation. The detailed information in foggy scenes can be improved by image enhancement. These methods do not need to consider the physical causes (such as fog/haze) of image degradation, and can effectively avoid the a priori estimation of the scene depth and depth mapping process. Representative defogging methods include: histogram equalization (Reza, 2004; Thomas et al., 2011), retinex-based methods (Rahman et al., 2004), homomorphic filter (Yu et al., 2015), wavelet transform (Rong and Jun, 2014; Jin et al., 2018a), and image fusion-based defogging methods (Li Y. et al., 2017; Galdran, 2018). These methods enhance both image contrast and saturation, so as to improve image visual quality. The detailed image information is first extracted from a single foggy image, and then fused to restore the details of the blurred areas. However, the defogging result obtained by the simply fusion of the two images cannot preserve all the detailed information of the scene in the original

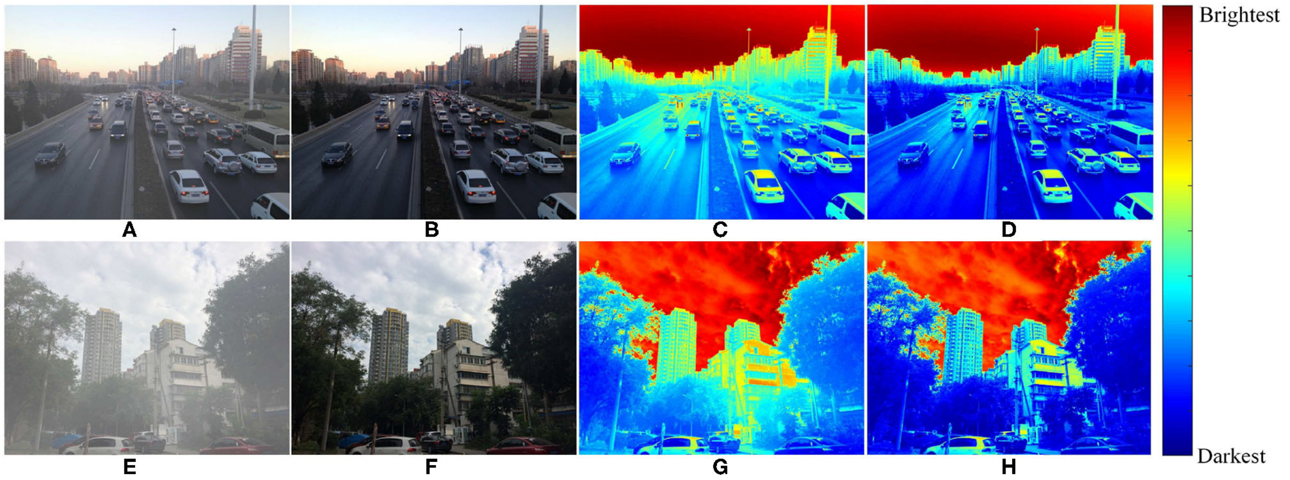


FIGURE 1 | Foggy and fog-free images and their heat maps. **(C,G)** Represent the heat maps of foggy images **(A,E)**, respectively. **(D,H)** Represent the heat maps of fog-free images **(B,F)**, respectively.

foggy image. To improve the detail preservation ability of image fusion techniques in the defogging process. Galdran (2018) introduced multi-exposure fusion techniques into image defogging. Multiple images with different exposure levels were extracted from one image by gamma correction, and saturation and contrast were considered as the weights of fusion. Multi-exposure fusion method was used to improve image visual quality from the global enhancement. However, some local information may be ignored in the global enhancement process, which affects the definition of the final output images. Therefore, it is necessary to optimize both global and local exposure, respectively (Qi et al., 2020).

To solve the above issues, this paper proposes a single image defogging method based on image patch decomposition and multi-exposure fusion. Since fog density is sensitive to contrast, gamma function is used to restore the details of local information by adjusting image contrast. A single input foggy image is corrected by gamma correction, so a set of underexposed images with different contrast are obtained. Spatial linear saturation enhancement is applied to the underexposed and original images, and then a set of foggy images with contrast and saturation enhancement are obtained. To retain more detailed information, images decomposition and fusion are used to enhance the detailed information of foggy images. With the help of a guided filter, each of multi-exposure images obtained after saturation adjustment is decomposed into the base and detail layers in the spatial domain. The guided filter does not damage any structure and detailed information of the processed images. In the base layer, a fixed-size moving window is used to extract image patches, and the best-quality areas are selected from each image patch for the fusion of image patches. According to the exposure features of each input image, the value of each pixel in the detail layer is estimated in the optimal exposure mode. The weight maps of both base and detail layers are constructed for image fusion. So, the fog-free image is obtained after fusing the base and detail layers. This paper has two main contributions as follows.

1. The proposed method can effectively avoid the complex process of both scene depth a priori estimation and depth mapping. A set of underexposed images are obtained by adjusting the contrast of foggy images. Spatial linear saturation adjustment is used to improve image saturation. Local features of foggy images are optimized by image patch structure decomposition to enhance the visual quality of fog-free images.
2. The proposed method can further improve the visual quality of the obtained fog-free images. Each exposure image is decomposed into based and detail layers. In the base layer, the local exposure quality is optimized by image patch structure decomposition. In the detail layer, the global exposure quality is optimized by the exposure degree evaluation model.

The rest of this paper is organized as follows. Section 2 discusses the related work; Section 3 elaborates the proposed solution in detail; Section 4 analyzes the comparative experimental results; and Section 5 concludes this paper.

2. RELATED WORK

Some researchers regard image defogging as a type of image restoration, so fog-free images can be obtained by an atmospheric light scattering model (Gonzalez et al., 2014). As a representative solution, dark channel prior (DCP) method proposed by He et al. (2009) makes at least one low-intensity pixel in a color channel of the local neighborhood around each pixel. This method learns the mapping relationship between a foggy image and the corresponding scene depth, and uses the value of the learned image transmission map to retrieve a physical model, so as to obtain the fog-free image by physical model calculation. Zhu et al. (2015) established a linear model based on the a priori information of a foggy image. According to the a priori scene depth information, an atmospheric scattering model is used to estimate transmittance and restore scene radiance, so as to effectively eliminate fog from a single image. He et al. (2016) proposed a convex optimization formula for image defogging. In the proposed foggy image model, bilinear coupled foggy images and light transmission distribution term are established to directly reconstruct the fog-free image. Fan et al. (2016) constructed a two-layer Gaussian process regression model, which established the relationship between an input image and its depth information transmission. In this method, the a priori knowledge of the local image structure is learned, and the multi-scale feature vectors of the input image are mapped to the corresponding transmitted light. The training model is used to restore the fog-free image. Wang et al. (2019) found that fuzzy regions were mainly concentrated on the luminance channel of YCrCb color space. So, the texture information lacking in the luminance channel can be recovered to enhance the visual contrast of foggy scenes. Yuan et al. (2017) introduced the gaussian mixture model (GMM). Based on haze density feature maps, an input foggy image is segmented into multiple scenes. The segmentation results can effectively identify sky areas that DCP cannot handle well. In the improved DCP model (Singh and Kumar, 2017), the atmospheric veil enhancement estimation is obtained by using the joint trilateral filter, and transmission maps are redefined to reduce the color distortion. Liu et al. (2017) proposed a ground radiation suppressed haze thickness map (GRS-HTM) based on haze thickness map (HTM) to calculate the fog distribution in the foggy image. The visible bands are affected by fog density. Fog components of each band are calculated by GRS-HTM to restore the fog-free image. Fog density changes with the depth of scene, so the degradation of image quality is also spatially variable. Atmospheric degradation model depends on the depth information of the corresponding scene, but the acquisition of scene depth information is unstable. This affects the accurate estimation of fog distribution and defogging performance. Without relying on the scene depth information, image enhancement-based defogging methods were proposed.

Image enhancement-based defogging methods mainly focus on enhancing both image contrast and saturation and highlighting image details. Yu et al. (2015) converted foggy images from RGB to HSV space. The overlapped sub-patch homomorphic filter is applied to the luminance components, and the processed image is converted back to RGB space to obtain the fog-free image. Kim et al. (2017) combined DCP, contrast constrained adaptive histogram equalization and discrete wavelet transform (CLAHE-DWT). First, the estimation of transfer function is improved in DCP. Then, image contrast and definition are improved by CLAHE-DWT, respectively. Finally, images processed by CLAHE-DWT are fused to generate the enhanced image. Galdran et al. (2015) proposed an enhanced variable image dehazing (EVID) method. This method enhances the local low pixels by adjusting the gray world hypothesis. Image colors are restored by controlling saturation, and image contrast between different channels is also improved. Image fusion is an important method used in image defogging, which can effectively improve the image contrast, detail information and so on (Jin et al., 2020; Liu et al., 2020). In the same scene, since the imaging equipment cannot focus different depth objects at the same time, so multi-focus image fusion technology is used to extract different focus areas from multiple images to synthesize a clear image (Jin et al., 2018b; Liu et al., 2019b). A fusion framework decomposes the source image into high- and low-pass subbands. The high-pass subbands are processed by a phase congruency-based fusion rule, and the low-pass subbands are processed by a local Laplacian energy-based fusion rule. The fused image is obtain by inversely transforming the processed high-pass and low-pass subbands. The fused image not only contains the enhanced detailed features, but also retains the structural information of the source image (Zhu et al., 2019). Li Y. et al. (2017) first used an adaptive color normalization method to correct color distortion images, and then enhanced the local details of both original and color corrected images. Dark channel, sharpness, and saliency features were taken as the weight maps for image fusion, and the pyramid fusion strategy was used to reconstruct images. Liu et al. (2019a) first transformed the speckle noise into additive noise by logarithmic transformation. Then, the local image blocks are matched by Gray theory, the approximate low-rank matrices grouped by the similar blocks of the reference patches is obtained. Wavelet transform is used to estimate the noise variance of the noisy image. Finally, weighted nuclear norm minimization is used to the denoised image. Gao et al. (2020) obtained a set of self-constructed images with different exposure levels by segmenting atmospheric light range. Therefore, an adaptive multi-exposure image fusion method based on scale invariant feature transform (SIFT) flow was proposed. On the basis of fusion, self-constructed images with different exposure levels are adaptively selected by using two-layer visual sense selectors. Galdran (2018) applied the multi-exposure image fusion method to image defogging. The global image exposure quality is enhanced to improve the image visual quality. This method enhances the global image features, but the enhancement of local features is uncertain, which affects the image quality. On the same basis, Zhu et al. (2021) also used gamma correction to obtain a set of images with different exposure. By analyzing the global and local exposure, the weight maps are constructed to guide the fusion process. The defogged image is obtained after saturation adjustment. Zheng et al. (2020) directly adjusted the saturation of underexposed images after gamma correction, and proposed a fusion method based on adaptive decomposition of image patches. The adaptive selection of image patch size is realized by fitting both texture entropy and image patch size. High weights are assigned to image patches with good visual quality for image fusion. Similar to this method, this paper also proposes an image patch based multi-exposure fusion method for image defogging. Image restoration is achieved through the optimization of both local and global exposure quality.

Now, deep learning is widely used in image defogging. Cai et al. (2016) first applied deep learning to image defogging and proposed DehazeNet. This paper used DehazeNet to estimate a medium transmission map in an atmospheric scattering model. A hazy image as input, and outputs its medium transmission map. Then, a haze-free image is recovered by atmospheric scattering model. And a novel nonlinear activation function is proposed, the quality of recovered haze-free image is improved by this function. Zhang and Patel (2018) proposed a new single image dehazing method, called densely connected pyramid dehazing network (DCPDN). DCPDN includes two generators, which are used to generate the transmission map and the atmospheric light, respectively. A new edge-preserving densely connected encoder-decoder structure with multi-level pyramid pooling module is designed to estimate the transmission map. Then the U-net structure is used to estimate the atmospheric light.Both the transmission map and the atmospheric light are introduced into an atmospheric scattering model to restore the fog-free image. A joint-discriminator based on generative adversarial network (GAN) framework is proposed to further incorporate the mutual structural information between the estimated transmission map and the dehazed result. This kind of defogging method using network estimation parameters still needs the help of atmospheric scattering model. Li B. et al. (2017) proposed an image dehazing model built with a CNN, called All-in-One Dehazing Network (AOD-Net). This paper dosed not estimate the transmission map and the atmospheric light separately, but directly generated clear images through light-weight CNN. Qin et al. (2020) proposed an end-to-end feature fusion attention network (FFA-Net) for single image dehazing. This paper combined channel attention and pixel attention mechanism to form a novel feature attention (FA) module. FA focused more attention on the thick haze pixels and more important channel information. And local residual learning allows the less important information to be bypassed through multiple skip connections. To giving more weight to important features, an attention-based different levels feature fusion (FFA) structure is proposed, the feature weights are adaptively learned from FA.

3. THE PROPOSED IMAGE DEFOGGING METHOD

As shown in **Figure 2**, the proposed single image defogging method performs gamma correction on an input foggy image to obtain a set of underexposed images. Both the underexposed images and the original image are enhanced by spatial linear saturation. All the images are decomposed into base and detail layers by a guided filter. A fixed-size moving window is used to

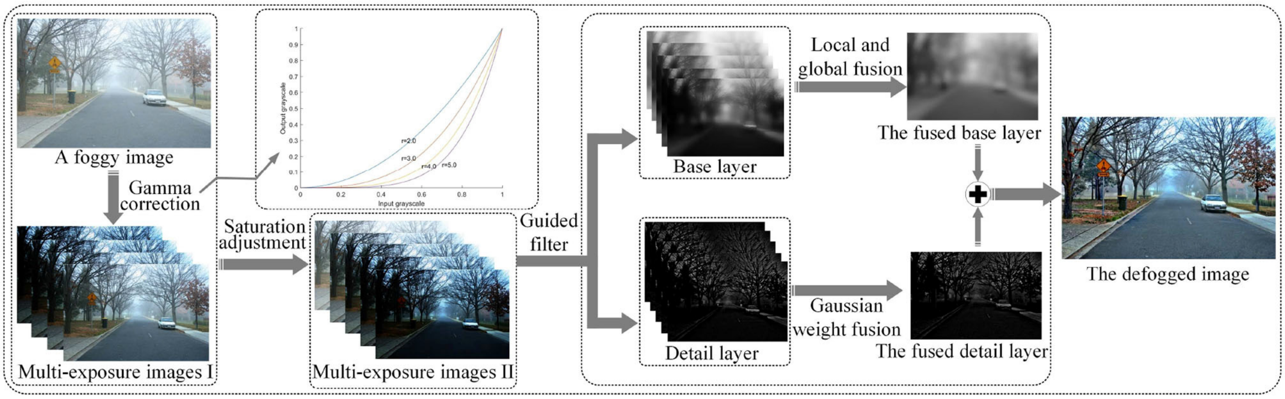


FIGURE 2 | The proposed image defogging framework based on image patch and multi-exposure fusion.

extract image patches from the base layer. Low-level features such as signal strength, signal structure, and mean intensity are used to improve fusion quality. Image patches are decomposed into signal strength, signal structure, and mean intensity by a structure decomposition method. The best-quality areas of the above three low-level features are selected for fusion. The whole luminance components of each input image are used to extract exposure features, and the extracted features are applied to optimize the global exposure quality of detail layer.

3.1. Image Preprocessing by Gamma Function

Gamma correction is used to adjust an input foggy image $I(x)$ nonlinearly by increasing or decreasing the exposure of the input image to change the local contrast of blurry areas.

$$I(x) \mapsto \alpha \cdot I(x)^{\gamma} \tag{1}$$

where α and γ are positive numbers. When $\gamma < 1$, the gray level of bright areas is compressed. The gray level of dark areas is stretched to be brighter, and the whole image becomes bright, which causes the color tone of high-luminance contents to be too bright. So, the detailed contents are not obvious in human visual perception (Galdran, 2018). On the contrary, when $\gamma > 1$, the whole image darkens and a series of underexposed images are obtained, and the image details are highlighted. For the input foggy image $I(x)$, the contrast Y of the given area Ω is shown as follows.

$$Y(\Omega) = y_{I^{\Omega}_{\max}} - y_{I^{\Omega}_{\min}} \tag{2}$$

where $y_{I^{\Omega}_{\max}} = \max\left\{y_{I(x)} \,|x \in \Omega\right\}$ and $y_{I^{\Omega}_{\min}} = \min\left\{y_{I(x)} \,|x \in \Omega\right\}$. When $\gamma > 1$, a set of underexposed images are obtained by Equation (2). Gamma correction is a kind of global correction, and the contrast of some areas with moderate exposure is reduced. As shown in **Figure 3**, the value of γ is 2, 3, 4, or 5, respectively, four foggy images with different exposure are obtained by gamma correction. Different exposure images highlight the details of different areas.

3.2. Saturation Enhancement

The input foggy image $I(x)$ is corrected by gamma ray to obtain a set of multi-exposure image sequences $Q = \{I_1(x), I_2(x), ..., I_N(x)\,|N = 5\}$. Each image has $I_n(x) = \left[I_n^R(x), I_n^G(x), I_n^B(x)\right]$. For each image, the maximum and minimum values of each pixel are calculated.

$$\begin{cases} rgb_{\max} = \max(R, \max(G, B)) \\ rgb_{\min} = \min(R, \min(G, B)) \end{cases} \tag{3}$$

When $\Delta = \left(rgb_{\max} - rgb_{\min}\right)/255 > 0$, the saturation P of each pixel in an image is calculated as follows.

$$P = \begin{cases} \Delta/value, L < 0.5 \\ \Delta/\left(2 - value\right), L \geq 0.5 \end{cases} \tag{4}$$

where $value = \left(rgb_{\max} + rgb_{\min}\right)/255$ and $L = value/2$. The saturation of each pixel is normalized. The same adjustment operation is performed on the three channels of RGB, and the adjustment of saturation increment for each image is within $[-100, 100]$.

When $Increment \geq 0$, the three channels of RGB are adjusted by Equation (5).

$$I'_n(x) = I_n(x) + [I_n(x) - L \times 255] \times \alpha \tag{5}$$

where $\alpha = 1/\beta - 1$ and $I'_n(x) = \left[I_n^{R'}(x), I_n^{G'}(x), I_n^{B'}(x)\right]$ represents the saturation of an image after saturation adjustment.

$$\beta = \begin{cases} P, & Increment + P \geq 1 \\ 1 - Increment, & \text{else} \end{cases} \tag{6}$$

When $Increment < 0$, the three channels of RGB are adjusted by Equation (7).

$$I'_n(x) = I_n(x) + [I_n(x) - L \times 255] \times (1 + \alpha) \tag{7}$$

where $\alpha = Increment$.

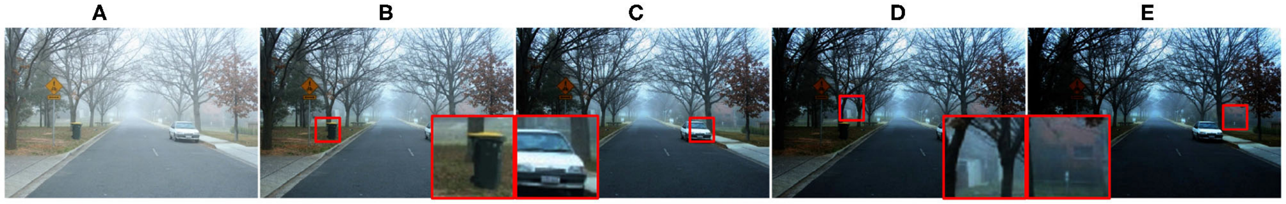


FIGURE 3 | Original image is corrected by gamma function. **(A)** A foggy image, **(B)** $\gamma = 2$, **(C)** $\gamma = 3$, **(D)** $\gamma = 4$, **(E)** $\gamma = 5$.

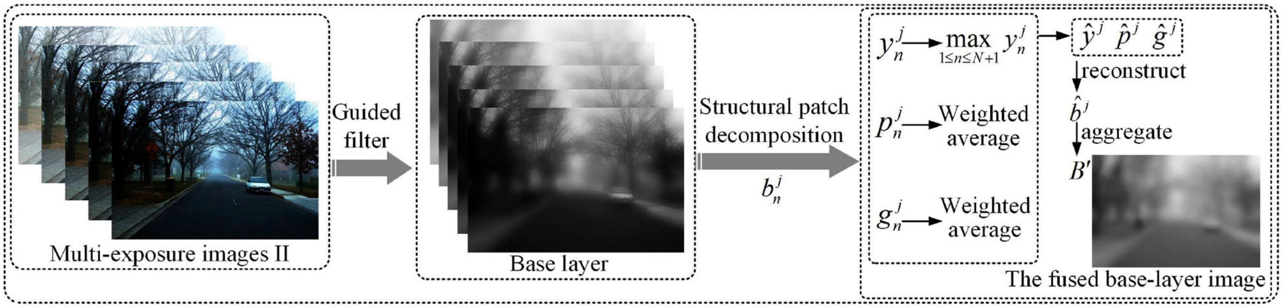


FIGURE 4 | The fusion process of the base layer. b_n^j represents image patches. y_n^j, p_n^j, and g_n^j represent signal strength, signal structure, and mean intensity, respectively. y^j, p^j, and g^j represent the desired signal strength, signal structure, and mean intensity, respectively. $\hat{b}^j$ represents the fusion of image patches, B' is the fused base-layer image.

3.3. Multi-Exposure Image Fusion Defogging

3.3.1. Image Decomposition by a Guided Filter

The input images $\left\{I_n'(x)\,|1 \le n \le N, N = 5\right\}$ is decomposed into the base and detail layers. Luminance component G_n of the input image is calculated by the weighted sum of the three channels of RGB. Since a guided filter can keep edge-preservation smooth (Li et al., 2012), the base layer is obtained by a guided filter as follows.

$$B_n = T_{r,\delta}\left(G_n, G_n\right) \tag{8}$$

where $T_{r,\delta}(Z, H)$ is a guided filter operator, r is the filter radius, and δ is used to control fuzzy degree. Z and H represent both input image and guide image, respectively. G_n represents both input image and guide image (Nejati et al., 2017). The detail layer D_n is obtained as follows.

$$D_n = I_n'(x) - B_n \tag{9}$$

3.3.2. Fusion Defogging Based on Global and Local Optimization

As shown in **Figure 4**, the optimization of both global and local exposure is realized by structure decomposition. A fixed-size moving window is used to extract image patches $b_n^j = \left\{b_n^j\,|1 \le n \le N, 1 \le j \le J\right\}$ from the base layer, b_n^j represents the j-th image patch of the n-th image. Structure decomposition proposed in Ma et al. (2017) is used to decompose image patches. Image patches are decomposed into three parts by Equation (10): signal strength y_n^j, signal structure p_n^j, and mean intensity g_n^j.

$$\begin{aligned} b_n^j &= \left\| b_n^j - \mu_{b_n^j} \right\| \cdot \frac{b_n^j - \mu_{b_n^j}}{\left\| b_n^j - \mu_{b_n^j} \right\|} + \mu_{b_n^j} \\ &= \left\| \tilde{b}_n^j \right\| \cdot \frac{\tilde{b}_n^j}{\left\| \tilde{b}_n^j \right\|} + \mu_{b_n^j} \\ &= y_n^j \cdot p_n^j + g_n^j \end{aligned} \tag{10}$$

where $\mu_{b_n^j}$ is the mean value of each image patch, and $\|\cdot\|$ is the l_2-norm of the vector.

The highest signal strength of all image patches at the same spatial position in the image set is taken as the expected signal strength $\hat{y}^j$ of the fused image patch.

$$\hat{y}^j = \max_{1 \le n \le N} y_n^j = \max_{1 \le n \le N} \left\| \tilde{b}_n^j \right\| \tag{11}$$

To obtain the expected image patch signal structure, the weighted average of the signal strength of input image patch set is calculated as follows.

$$\hat{p}^j = \frac{\sum_{n=1}^{N} P\left(\tilde{b}_n^j\right) p_n^j \Big/ \sum_{n=1}^{N} P\left(\tilde{b}_n^j\right)}{\left\| \sum_{n=1}^{N} P\left(\tilde{b}_n^j\right) p_n^j \Big/ \sum_{n=1}^{N} P\left(\tilde{b}_n^j\right) \right\|} \tag{12}$$

where the weight function $P\left(\tilde{b}_n^j\right) = \left\| \tilde{b}_n^j \right\|^t$ determines the contribution of each image patch to the fused image patch, and

$t \geq 0$ is an exponential parameter. When the value of t gets larger, the image patch with higher intensity is highlighted.

The exposure quality of each image patch in the input image is measured by a two-dimensional gaussian function.

$$G\left(\mu_n, g_n^j\right) = \exp\left(-\frac{(\mu_n - 0.5)^2}{2\delta_\mu^2} - \frac{\left(g_n^j - 0.5\right)^2}{2\delta_g^2}\right) \tag{13}$$

where δ_μ and δ_g are the gaussian standard deviations of the constructed two-dimensional gaussian function. δ_μ and δ_g control the expansion of contour along μ_n and size g_n^j, respectively. The expected mean intensity $\hat{g}^j$ of the image patch is shown as follows.

$$\hat{g}^j = \frac{\sum_{n=1}^{N} G\left(\mu_n, g_n^j\right) g_n^j}{\sum_{n=1}^{N} G\left(\mu_n, g_n^j\right)} \tag{14}$$

$\hat{y}^j$, $\hat{p}^j$, and $\hat{g}^j$ form a new vector. The fused image patch $\hat{b}^j$ is represented as follows.

$$\hat{b}^j = \hat{y}^j \cdot \hat{p}^j + \hat{g}^j \tag{15}$$

To optimize the local-exposure quality, a fixed-size moving window is used to extract image patches at the same spatial position from the base layer of the input image. The pixels in the overlapped image patches are averaged. The above steps of the decomposition and fusion of image patches are repeated, and then $\sum_{j=1}^{J} \hat{b}^j$ is used to obtain the fused image B' of the base layer.

Two dimensional gaussian function is used to evaluate the exposure quality of B' and optimize the global exposure quality of B'. The mixed weight $\mathrm{E}_{n,B}$ of each pixel (x, y) in B'_n is calculated as follows.

$$\mathrm{E}_{n,B}\left(x, y\right) = \exp\left(-\frac{\left(B'\left(x, y\right) - 0.5\right)^2}{2\delta_\mu^2} - \frac{\left(\bar{G} - 0.5\right)^2}{2\delta_g^2}\right) \tag{16}$$

$\hat{B}$ represents the weighted sum of each base-layer image in the input image set and its corresponding weight $\mathrm{E}_{n,B}$ in the fused image.

$$\hat{B} = \sum_{n=1}^{N} \mathrm{E}_{n,B} B' \tag{17}$$

3.3.3. Exposure Fusion Image Based on Gaussian Weight Method

Each luminance component is convoluted with a 7 × 7 average filter to simply calculate the exposure features $\varphi_n\left(x, y\right)$ of each pixel in multi-exposure image set, and $\varphi_n\left(x, y\right)$ is the mean intensity of a small area around the pixel (x, y). The value of each pixel in the detail layer in the optimal exposure mode is estimated by analyzing the shading changes of different pixels. The weight $\mathrm{E}_{n,D}\left(x, y\right)$ of each pixel (x, y) in the detail layer of the n-th input image is calculated by using the exposure degree evaluation model.

$$\mathrm{E}_{n,D}\left(x, y\right) = \exp\left(-\frac{\left(\varphi_n\left(x, y\right) - \varphi_0\right)^2}{2\delta_d^2}\right) \tag{18}$$

where $\varphi_n\left(\cdot\right)$ is the exposure feature, δ_d is the gaussian standard deviation, and φ_0 as the best exposure constant equals the middle value of the intensity range.

The defogged image is defined as follows.

$$J\left(x\right) = \hat{B} + \omega \sum_{n=1}^{N} \mathrm{E}_{n,D} D_n \tag{19}$$

where $\omega \geq 1$ controls the detail intensity and local contrast of the defogged image $J\left(x\right)$. According to the experimental results of the fusion performance, the value of ω is set to 1.1.

3.3.4. Verification of Image Intensity Reduction After Defogging

Koshmieder proposed an atmospheric scattering model to solve the image degradation issues caused by fog (Gonzalez et al., 2014).

$$I\left(x\right) = t\left(x\right) J\left(x\right) + A\left(1 - t\left(x\right)\right) \tag{20}$$

where $I\left(x\right)$ represents a foggy image. $J\left(x\right)$ represents the corresponding fog-free image of $I\left(x\right)$. A represents the global atmospheric light. $t\left(x\right)$ is the transmitted light. $t\left(x\right) J\left(x\right)$ describes the radiation and attenuation of the scene in the medium. $A\left[1 - t\left(x\right)\right]$ is the atmospheric light formula.

Equation (20) that reduces image intensity is used to formalize foggy images. In this paper, underexposure or overexposure processing is applied to foggy images, and the corresponding exposure results are fused to obtain the image areas with good exposure quality. To meet the requirement of image intensity reduction, the proposed method is only applied to the underexposed images to reduce global exposure. When $\gamma > 1$, it is easy to verify that the fused image obtained by using $B'_n = \sum_{j=1}^{J} \hat{b}_n^j$ always meets the requirement of image intensity reduction.

Proof:

In Zheng et al. (2020), it simply verifies that the fusion of the images obtained after gamma correction, saturation linear adjustment and image structure decomposition meets the requirement of intensity reduction $J\left(x\right) \leq I\left(x\right)$. The proof is shown as follows.

Given a set of gamma parameters $\Gamma = \left\{\gamma^1, \gamma^2, ..., \gamma^K | \gamma^k > 1\right\}$, a set of underexposed images $Q = \left\{I_1\left(x\right), I_2\left(x\right), ..., I_{N-1}\left(x\right)\right\}$ is obtained. Since $I\left(x\right) \in [0, 1]$, $I(x)^{\gamma^k} < I\left(x\right)$ is available for all pixels. Due to the invariance principle of brightness in the linear adjustment of saturation, the pixel intensity component is $I\left(x\right) = \frac{1}{3}\left(R + G + B\right)$ (Gonzalez and Woods, 1977). Therefore, for any foggy image, $I\left(x\right) = Q'_n\left(x\right)$ is satisfied before and after saturation adjustment. Therefore, all the pixels after saturation adjustment satisfy $\left(Q_n(x)^{\gamma^k}\right)' < I\left(x\right)$.

Since an image patch $b_n^j \in I(x)$, all $b_n^j(x)^{\gamma^k} \in \left(Q_n(x)^{\gamma^k}\right)'$ satisfy $b_n^j(x)^{\gamma^k} < b_n^j(x)$. Therefore, image patches can meet the requirements of image intensity reduction after gamma correction and saturation adjustment.

According to the above proof $b_n^j \in I(x)$ is satisfied for any image patch. The structure decomposition of image patches is performed on both sides of Equation (21) (Ma et al., 2017).

$$\left(\left(y_n^j\right)^{\gamma^k} \cdot \left(p_n^j\right)^{\gamma^k} + \left(g_n^j\right)^{\gamma^k}\right) < \left(y_n^j \cdot p_n^j + g_n^j\right) \tag{21}$$

Since y_n^j, p_n^j, and g_n^j of each image patch are unit length vectors, and the initial foggy image $I(x)$ is the input of image fusion. Therefore, the expected contrast of the fused image patch satisfies $\hat{y}^j = \max\limits_{1 \leq n \leq N+1} y_n^j \leq y^j$. Similarly, since the weight of the mean luminance is $\sum_{n=1}^{N}\left(\frac{G\left(\mu_n, g_n^j\right)}{\sum_{n=1}^{N} G\left(\mu_n, g_n^j\right)}\right) = 1$, the expected average brightness is satisfied as follows.

$$\hat{g}^j = \frac{\sum_{n=1}^{N} G\left(\mu_n, g_n^j\right) g_n^j}{\sum_{n=1}^{N} G\left(\mu_n, g_n^j\right)} < g^j \tag{22}$$

The mode of signal structure satisfies $\left\|p_n^j\right\| = \left\|p^j\right\|$. So, $\hat{b}^j = \hat{y}^j \cdot \hat{p}^j + \hat{g}^j \leq b_n^j$. Image patches meet the requirements of image intensity reduction after structural decomposition. Since $\hat{b}^j \in J(x)$ follows $\hat{b}^j \leq b_n^j$, $J(x) \leq I(x)$. So, the fused image always meets the requirements of image intensity reduction.

4. EXPERIMENTAL ANALYSIS

4.1. Experiment Preparations

Eighty three real-world foggy natural images with different sizes are used in the comparative experiments. These images can be downloaded from http://live.ece.utexas.edu/research/fog/fade_defade.html, http://github.com/agaldran/amef_dehazing, http://github.com/JiamingMai/Color-Attenuation-Prior-Dehazing or captured by ourselves. A synthetic foggy image dataset (RESIDE) with 100 scene images (Li et al., 2019) downloaded from http://sites.google.com/view/reside-dehaze-datasets. One hundred remote-sensing geographic images were collected from Google Earth by ourselves. Seventeen real-world tunnel images were collected by ourselves. Thirteen image defogging methods are used for comparison, which are AMEF (Galdran, 2018), CAP (Zhu et al., 2015), CO (He et al., 2016), DCP (He et al., 2009), DEFADE (Choi et al., 2015), GPR (Fan et al., 2016), MAMF (Cho et al., 2018), OTE (Ling et al., 2018), WCD (Chiang and Chen, 2012), DehazeNet (Cai et al., 2016), FFA-Net (Qin et al., 2020), a novel fast single image dehazing algorithm based on artificial multiexposure image fusion (MIF) (Zhu et al., 2021) and the proposed defogging method. All the experiments were programmed by MATLAB 2016b and run on a desktop with an Intel I9-7900X@3.30 GHz CPU and 16.00 GB RAM.

4.2. Subjective Visual Evaluation

As shown in **Figures 5–9**. The results of five different scenes are selected to confirm that the proposed method has good defogging performance.

Figure 5 compares the defogging performance of thirteen methods on a real-world natural image. As shown in **Figures 5C–K**, the performance of CAP, CO, GPR, DehazeNet is poor. In the magnified areas, the details of the mountain are not visible. The hues shown in **Figures 5E,I** deviate. The global brightness of DEFADE and WCD as shown in **Figures 5F,J** respectively is low, and the fog shown in the magnified areas of **Figure 5J** is not completely removed. The brightness and saturation of **Figure 5L** are low. Although MAMF restores the high saturation of the source image, the contrast is sacrificed in the defogged image shown in **Figure 5H**, and the loss of structural and texture details can be seen from the magnified areas. As shown in **Figures 5B–N**, compared with other 10 methods, AMEF, MIF, and the proposed method achieve better defogging performance in local details and global brightness. The global saturation of the defogged image obtained by MIF or the proposed method is slightly better than the one obtained by AMEF.

Figure 6A is a real-world rural natural image. Due to the poor defogging performance of DCP and OTE, the color of sky is distorted, and the details shown in the magnified areas are lost, as shown in **Figures 6E,I**. In **Figures 6D–L**, the overall brightness of defogged images is too low, and the details shown in the magnified areas are lost. CAP and WCD have poor defogging performance. As shown in **Figure 6C**, there is no obvious change after defogging. The image saturation of **Figure 6J** is too low. As shown in **Figures 6B**, **7C–N**, the image visibility is greatly improved, and the details shown in the magnified areas are clear. However, color distortion appears in the sky of **Figure 6H**. AMEF, MIF, and the proposed method have the best image defogging performance. The comparative results show that the overall brightness of the defogged image obtained by MIF or the proposed method is slightly better than the one obtained by AMEF.

Figure 7 compares the defogging performance of thirteen methods on a synthetic driving image. As shown in **Figures 7E,I**, the color of some areas in images is distorted, and the details shown in the magnified areas are lost. GPR have poor defogging performance, the clarity of the image decreased after defogging, as shown in **Figure 7G**. As shown in **Figures 7C–K**, the overall brightness of defogged images is too low, and the details shown in the magnified areas are lost. The sharpening degree of MAMF is too much, as shown in **Figure 7H**. In **Figures 7F,L**, some details information shown in the magnified areas are lost. As shown in **Figures 7B–N**, compared with other 10 methods, AMEF, MIF, and the proposed method have the best image defogging performance. The saturation of MIF and the proposed method is closer to the human eye observation habits than AMEF.

Figure 8 compares the defogging performance of 13 methods on a remote-sensing geographic image. As shown in **Figures 8D–F**, the details of the magnified areas are missing. The overall blurring degree of the defogged image obtained by GPR increases. The saturation of **Figures 8H,I** is too high,

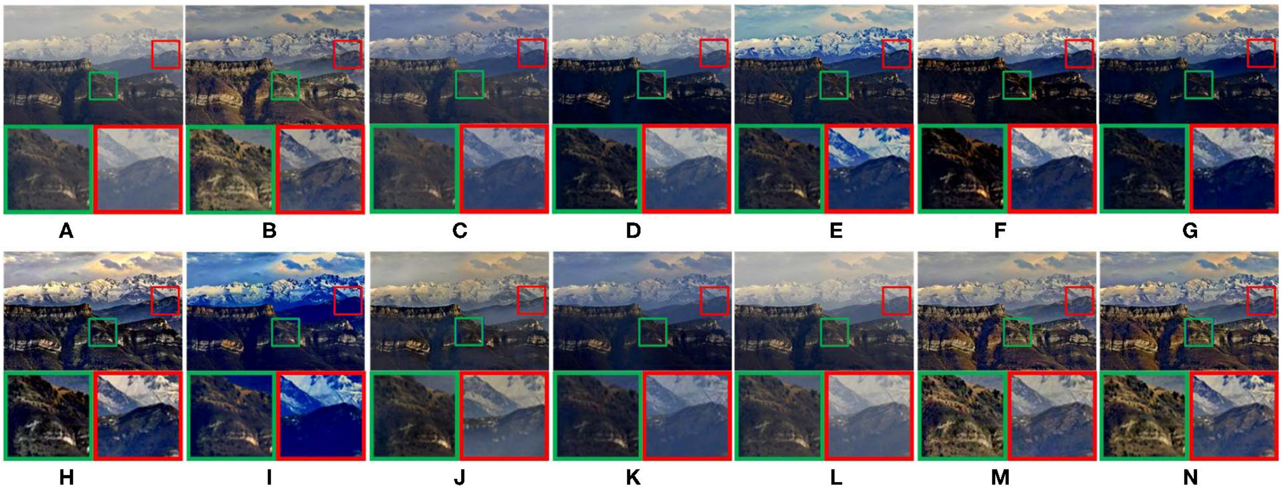


FIGURE 5 | Real-world natural image. **(A)** Represents the original foggy image. **(B–N)** Represent foggy-free images processed by AMEF, CAP, CO, DCP, DEFADE, GPR, MAMF, OTE, WCD, DehazeNet, FFA-Net, MIF, and the proposed method.

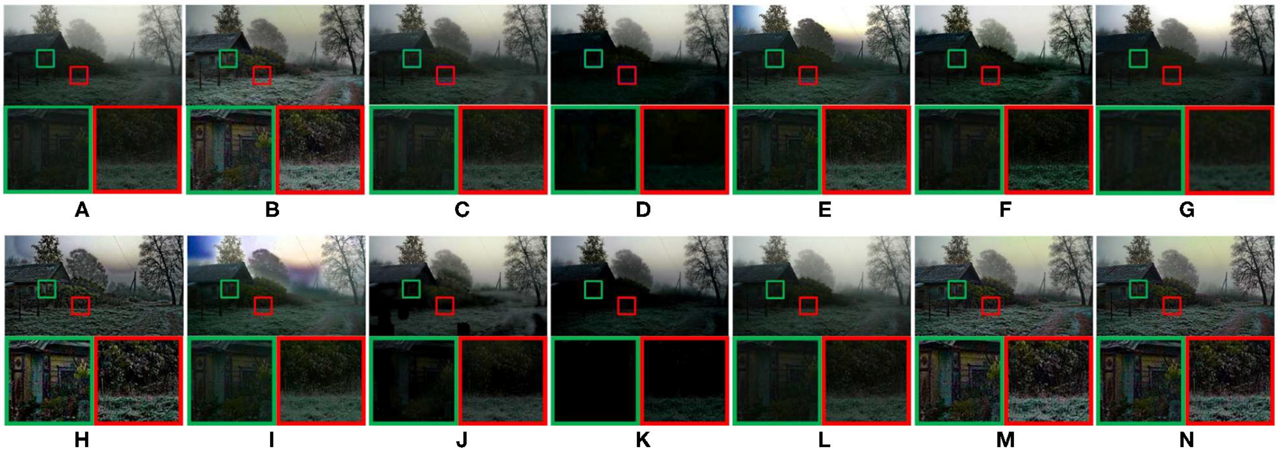


FIGURE 6 | Real-world rural natural image. **(A)** Represents the original foggy image. **(B–N)** Represent foggy-free images processed by AMEF, CAP, CO, DCP, DEFADE, GPR, MAMF, OTE, WCD, DehazeNet, FFA-Net, MIF, and the proposed method.

which leads to color distortion. The details of the magnified areas of **Figure 8I** are lost. As shown in **Figure 8J**, there is obvious contrast between light and dark light in the magnified areas. **Figures 8B,K–N** show good defogging performance, the overall brightness of the defogged images is good. However, the details shown in the magnified areas are lost, as shown in **Figures 8K,L**. After removing fog from the remote-sensing geographic image, it is helpful to recognize the objects shown in the remote-sensing geographic images and improve the recognition accuracy.

Figure 9 shows the defogged tunnel images obtained by 13 methods. The defogged image obtained by OTE has high saturation and color distortion, as shown in **Figure 9I**. In **Figures 9C–L**, obvious fog residue exists. The defogged image obtained by WCD has obvious distortion, as shown in **Figure 9J**. The overall brightness of **Figure 9E** is low. The overall brightness of **Figures 9G,H** is high, and the saturation is low. The saturation of **Figure 9M** is high. DEFADE, AMEF, DehazeNet, and the proposed method achieve good defogging performance. As shown in the magnified areas of **Figure 9F**, high saturation can reduce image contrast, and the texture details of tunnel wall are lost. After defogging tunnel images, the cracks on the inner wall of the tunnel and the pavement damages are well-observed.

4.3. Objective Evaluation

Structural similarity (SSIM) (Wang et al., 2004), peak-signal-to-noise ratio (PSNR) (Hore and Ziou, 2010), fog aware density evaluator (FADE) (Choi et al., 2015), and Entropy (Qing et al., 2016) are used as objective evaluation indexes. SSIM is used to measure the similarity between the defogged and reference images. The high SSIM value means the high similarity between the foggy and defogged images. PSNR is used to measure the distortion of defogging image compared with reference image.

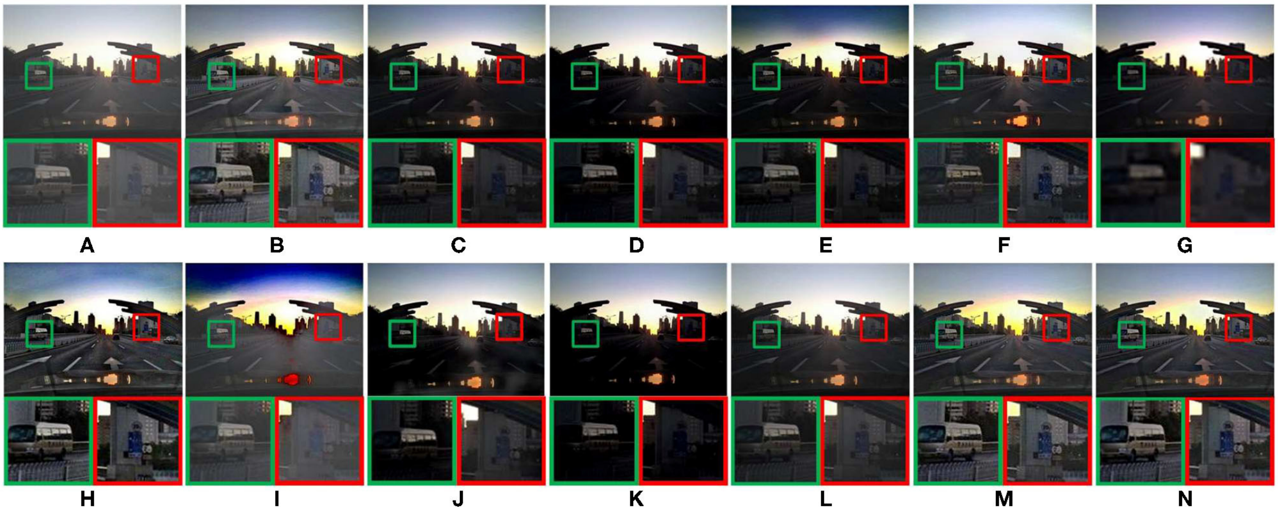


FIGURE 7 | Synthetic driving image. **(A)** represents the original foggy image. **(B–N)** Represent foggy-free images processed by AMEF, CAP, CO, DCP, DEFADE, GPR, MAMF, OTE, WCD, DehazeNet, FFA-Net, MIF, and the proposed method.

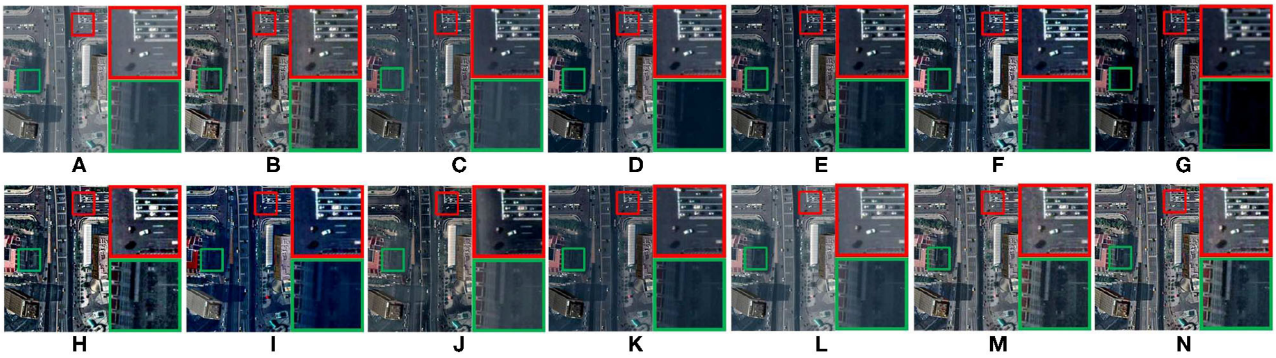


FIGURE 8 | Remote-sensing geographic image. **(A)** Represents the original foggy image. **(B–N)** Represent foggy-free images processed by AMEF, CAP, CO, DCP, DEFADE, GPR, MAMF, OTE, WCD, DehazeNet, FFA-Net, MIF, and the proposed method.

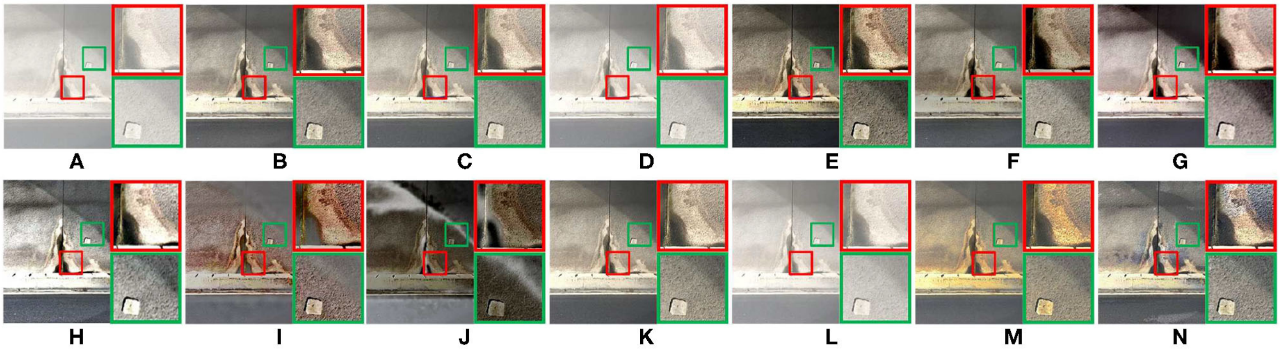


FIGURE 9 | Tunnel image. **(A)** Represents the original foggy image. **(B–N)** Represent foggy-free images processed by AMEF, CAP, CO, DCP, DEFADE, GPR, MAMF, OTE, WCD, DehazeNet, FFA-Net, MIF, and the proposed method.

The high PSNR value means less distortion of defogging image. FADE is a no-reference evaluation index of image defogging performance. The image blurring degree is directly proportional to the value of FADE. Entropy reflects the average amount of information in the image. A large Entropy value means the large average amount of information is retained. Thirteen defogging

TABLE 1 | Evaluation of two objective indexes in the real-world natural image (**Figure 5**) defogging experiment.

	AMEF	CAP	CO	DCP	DEFADE	GPR	MAMF	OTE	WCD	DehazeNet	FFA-Net	MIF	Proposed
FADE	0.4177	0.7179	0.5151	0.3797	0.3121(4)	0.4086	0.2055(1)	0.2885(3)	0.5986	0.5366	0.8632	0.3414	0.2685(2)
Entropy	7.0971	6.7348	6.3071	7.0064	6.9570	6.7810	7.5853(1)	6.5808	7.0994(4)	6.9760	6.9084	7.3268(3)	7.3465(2)

TABLE 2 | Evaluation of two objective indexes in the real-world rural natural image (**Figure 6**) defogging experiment.

	AMEF	CAP	CO	DCP	DEFADE	GPR	MAMF	OTE	WCD	DehazeNet	FFA-Net	MIF	Proposed
FADE	0.4169	0.7189	0.6898	0.4854	0.3776	0.3087(3)	0.1874(1)	0.5654	0.4919	0.3584	0.7282	0.3158(4)	0.2691(2)
Entropy	7.4687	7.2995	6.2526	6.3164	7.1563	7.3750	7.4412	7.5312(2)	7.1124	6.5062	7.5052(4)	7.5942(1)	7.5176(3)

TABLE 3 | Evaluation of two objective indexes in the synthetic driving image (**Figure 7**) defogging experiment.

	AMEF	CAP	CO	DCP	DEFADE	GPR	MAMF	OTE	WCD	DehazeNet	FFA-Net	MIF	Proposed
SSIM	0.8037	0.8904(3)	0.6737	0.7191	0.9273(2)	0.8221	0.7415	0.6733	0.5641	0.4868	0.9897(1)	0.8603	0.8645(4)
PSNR	29.198	25.983	26.114	24.650	33.323(3)	26.223	28.103	28.718	25.790	63.748(1)	37.461(2)	28.513	29.428(4)

TABLE 4 | Evaluation of two objective indexes in the remote-sensing geographic image (**Figure 8**) defogging experiment.

	AMEF	CAP	CO	DCP	DEFADE	GPR	MAMF	OTE	WCD	DehazeNet	FFA-Net	MIF	Proposed
FADE	0.4201	0.6580	0.4681	0.3367	0.3028(4)	0.3852	0.1915(2)	0.2479(3)	0.4045	0.5049	0.7200	0.4105	0.1907(1)
Entropy	7.3273(3)	6.4009	6.6120	6.8067	7.2608	6.5937	7.4313(2)	6.5041	7.0765	6.7150	7.0420	7.3230(4)	7.5685(1)

methods are applied to 300 foggy images. Five defogged images are selected for illustration.

As shown in **Table 1**. According to the FADE and Entropy indexes of MAMF, MAMF can effectively reduce the fog density and retain the image information as much as possible. The Entropy of MIF and WCD is high, but FADE index of MIF and WCD reflects that MIF and WCD cannot effectively reduce the fog density. The FADE score is high, and the defogging performance is not effective enough. OTE and DEFADE can effectively reduce the fog density, but the Entropy of OTE and DEFADE rank low. In the defogging process, OTE and DEFADE lose some image information. The results of FADE and Entropy show that the proposed method can achieve good defogging performance.

In **Table 2**, FADE index of GPR and MAMF reflect that GPR and MAMF can effectively reduce the fog density, but Entropy index is low, some image information is lost in the defogging process. Entropy scores of FFA-Net and OTE are high, but their FADE indexes reflect that the defogging performance of FFA-Net and OTE are not good enough. MIF and the proposed method achieve a good ranking in FADE and Entropy indexes. MIF and the proposed method can effectively reduce the fog density and retain more image information.

As shown in **Table 3**, CAP, DEFADE, FFA-Net, and the proposed method have the highest four scores in SSIM index, which means that defogged result can effectively retain the structural information of the original image. However, PSNR index of CAP is low which means that there is more distortion in the defogging image. The PSNR of DehazeNet is high, but SSIM index of DehazeNet reflects that the structural information of the original image cannot be effectively preserved. DEFADE, FFA-Net and the proposed method achieve a good ranking in SSIM and PSNR indexes. DEFADE, FFA-Net, and the proposed method can effectively retain the structural information of the original image and reduce image distortion.

As shown in **Table 4**, the Entropy index of AMEF and MIF reflects that AMEF and MIF can retain more image information in the process of defogging. But the FADE index ranking of AMEF and MIF is low, which proves that its defogging performance is poor. FADE index of OTE and DEFADE show that OTE and DEFADE can effectively reduce fog, but the score of Entropy is low. In the process of defogging, OTE and DEFADE lose some image information. MAMF and the proposed method achieve good results in FADE and Entropy. MAMF and the proposed method can ensure the high defogging performance and reduce the information loss during the defogging process.

According to FADE index in **Table 5**, DCP, OTE, WCD, and the proposed method can effectively reduce the fog density. However, the ranking of Entropy index of OTE and WCD show that more image information is lost in the defogging process. Entropy index of GPR and DehazeNet reflect that GPR and DehazeNet can retain most of image information in the defogging process, but the ranking of FADE index of GPR and DehazeNet is low. For DCP and the proposed method, their FADE and Entropy index rankings are high, which proves that

TABLE 5 | Evaluation of two objective indexes in the tunnel image (**Figure 9**) defogging experiment.

	AMEF	CAP	CO	DCP	DEFADE	GPR	MAMF	OTE	WCD	DehazeNet	FFA-Net	MIF	Proposed
FADE	0.9207	1.1100	1.8038	0.5151(2)	0.6831	1.0563	1.0287	0.5799(4)	0.4712(1)	1.1700	1.5285	0.8143	0.5566(3)
Entropy	7.2576	7.3430	6.6227	7.5081(2)	7.1100	7.6589(1)	7.3797	6.8962	7.3096	7.4397(4)	6.9371	7.0303	7.4451(3)

they achieve good defogging performance and can effectively retain image information.

The proposed method is more in line with human eye observation habits in color saturation, image brightness, and sharpness. The image details are effectively restored. In general, compared with the other 12 methods, the proposed method can achieve good defogging performance, reduce image distortion, and retain rich image information. For 300 foggy images, the average running time of AMEF, CAP, CO, DCP, DEFADE, GPR, MAMF, OTE, WCD, DehazeNet, FFA-Net, MIF, and the proposed methods were 2.8274, 3.1197, 6.1310, 3.4911, 85.7802, 433.5796, 3.9043, 38.7347, 7.3273, 7.6966, 302.5901, 1.8056, and 20.7910 s, respectively. Although the proposed method has good defogging performance and is widely used in various image scenes, the average processing time is relatively long owing to the high computational complexity.

5. CONCLUSION

The proposed method can effectively achieve fog removal without any a priori knowledge of the scene depth information. A single foggy image is first corrected by gamma correction, and then a set of underexposed images is obtained. Multi-exposure image set is composed of these underexposure images and the original foggy image. Next, the saturation of multi-exposure images is adjusted. The multi-exposure images are decomposed into the base and detail layers by a guided filter. The image details are enhanced by image patch decomposition. Low-level features such as mean intensity, signal strength, and signal structure are used to improve fusion quality. The best-quality areas are collected from each base-layer image patch for the fusion of image patches. The global exposure quality of the detail layer is optimized by using the global luminance components of each input image. The comparative experimental results confirm the effectiveness of the proposed method and its superiority over the state-of-the-art methods. The proposed method can be applied to natural images, synthetic images, remote-sensing geographic images, and tunnel images to improve image quality. This method includes image scale decomposition, exposure quality detection, base-layer image fusion, and detail-layer image fusion. These calculation processes can achieve effective image defogging, but also increase the computational complexity. In future, a simpler and more effective fusion strategy will be designed to reduce the calculation steps and the running time of image defogging, while maintaining defogging performance.

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/s.

AUTHOR CONTRIBUTIONS

QL did data curation, formal analysis, conceptualization, formulated methodology, funding acquisition, software, visualization, and writing—original draft preparation. YL did formal analysis, investigation, devised the methodology, visualization, validation, and writing—review and editing. KL did investigation, resources, supervision, data curation, devised the methodology, and reviewed and edited the manuscript. WL did investigation, project administration, supervision, resources, and data curation. YC did supervision, resources, project administration, and funding acquisition. HD did formal analysis, conceptualization, resources, and data curation. XJ did supervision, data curation, investigation, project administration, and conceptualization. All authors contributed to this paper and approved the submitted version.

FUNDING

This work was jointly supported by the National Key Research and Development Program of China (Nos. 2019YFB1600700 and 2017YFC08060010), The Key Technology Innovation Special of Key Industries of the Chongqing Science and Technology under Grant (Nos. cstc2019jscx-fxydX0017, cstc2019jscx-gksbX0071, and cstc2019jscx-gksbX0008).

REFERENCES

Cai, B., Xu, X., Jia, K., Qing, C., and Tao, D. (2016). Dehazenet: an end-to-end system for single image haze removal. *IEEE Trans. Image Process.* 25, 5187–5198. doi: 10.1109/TIP.2016.2598681

Chiang, J. Y., and Chen, Y.-C. (2012). Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* 21, 1756–1769. doi: 10.1109/TIP.2011.2179666

Cho, Y., Jeong, J., and Kim, A. (2018). Model-assisted multiband fusion for single image enhancement and applications to robot vision. *IEEE Robot. Autom. Lett.* 3, 2822–2829. doi: 10.1109/LRA.2018.2843127

Choi, L. K., You, J., and Bovik, A. C. (2015). Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Trans. Image Process.* 24, 3888–3901. doi: 10.1109/TIP.2015.2456502

Fan, X., Wang, Y., Tang, X., Gao, R., and Luo, Z. (2016). Two-layer gaussian process regression with example selection for image dehazing. *IEEE*

Trans. Circ. Syst. Video Technol. 27, 2505–2517. doi: 10.1109/TCSVT.2016.2592328

Fattal, R. (2008). Single image dehazing. *ACM Trans. Graph.* 27, 1–9. doi: 10.1145/1360612.1360671

Galdran, A. (2018). Image dehazing by artificial multiple-exposure image fusion. *Signal Process.* 149, 135–147. doi: 10.1016/j.sigpro.2018.03.008

Galdran, A., Vazquez-Corral, J., Pardo, D., and Bertalmio, M. (2015). Enhanced variational image dehazing. *SIAM J. Imag. Sci.* 8, 1519–1546. doi: 10.1137/15M1008889

Gao, Y., Su, Y., Li, Q., Li, H., and Li, J. (2020). Single image dehazing via self-constructing image fusion. *Signal Process.* 167:107284. doi: 10.1016/j.sigpro.2019.107284

Gonzalez, A., Vazquez-Corral, J., Pardo, D., and Bertalmío, M. (2014). A variational framework for single image dehazing. *Eur. Conf. Comput. Vis.* 8927, 259–270.

Gonzalez, R. C., and Woods, R. E. (1977). Digital image processing. *Prent. Hall Int.* 28, 484–486. doi: 10.1109/TASSP.1980.1163437

He, J., Zhang, C., Yang, R., and Zhu, K. (2016). "Convex optimization for fast image dehazing," in *2016 IEEE International Conference on Image Processing (ICIP)* (Phoenix, AZ), 2246–2250. doi: 10.1109/ICIP.2016.7532758

He, K., Sun, J., and Tang, X. (2009). "Single image haze removal using dark channel prior," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1956–1963.

Hore, A., and Ziou, D. (2010). "Image quality metrics: PSNR vs. SSIM," in *2010 20th International Conference on Pattern Recognition* (Istanbul), 2366–2369. doi: 10.1109/ICPR.2010.579

Jin, X., Chen, G., Hou, J., Jiang, Q., Zhou, D., and Yao, S. (2018a). Multimodal sensor medical image fusion based on nonsubsampled shearlet transform and S-PCNNS in HSV space. *Signal Process.* 153, 379–395. doi: 10.1016/j.sigpro.2018.08.002

Jin, X., Jiang, Q., Chu, X., Lang, X., Yao, S., Li, K., et al. (2020). Brain medical image fusion using L2-norm-based features and fuzzy-weighted measurements in 2-d littlewood-paley ewt domain. *IEEE Trans. Instrum. Meas.* 69, 5900–5913. doi: 10.1109/TIM.2019.2962849

Jin, X., Zhou, D., Yao, S., Nie, R., Jiang, Q., He, K., et al. (2018b). Multi-focus image fusion method using s-pcnn optimized by particle swarm optimization. *Soft Comput.* 22, 6395–6407. doi: 10.1007/s00500-017-2694-4

Ju, M., Ding, C., Zhang, D., and Guo, Y. J. (2019). BDPK: Bayesian dehazing using prior knowledge. *IEEE Trans. Circ. Syst. Video Technol.* 29, 2349–2362. doi: 10.1109/TCSVT.2018.2869594

Kim, K., Kim, S., and Kim, K.-S. (2017). Effective image enhancement techniques for fog-affected indoor and outdoor images. *IET Image Process.* 12, 465–471. doi: 10.1049/iet-ipr.2016.0819

Li, B., Peng, X., Wang, Z., Xu, J., and Feng, D. (2017). "AOD-net: All-in-one dehazing network," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice), 4780–4788. doi: 10.1109/ICCV.2017.511

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., et al. (2019). Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* 28, 492–505. doi: 10.1109/TIP.2018.2867951

Li, Y., Miao, Q., Liu, R., Song, J., Quan, Y., and Huang, Y. (2017). A multi-scale fusion scheme based on haze-relevant features for single image dehazing. *Neurocomputing* 283, 73–86. doi: 10.1016/j.neucom.2017.12.046

Li, Z. G., Zheng, J. H., and Rahardja, S. (2012). Detail-enhanced exposure fusion. *IEEE Trans. Image Process.* 21, 4672–4676. doi: 10.1109/TIP.2012.2207396

Ling, Z., Gong, J., Fan, G., and Lu, X. (2018). Optimal transmission estimation via fog density perception for efficient single image defogging. *IEEE Trans. Multimed.* 20, 1699–1711. doi: 10.1109/TMM.2017.2778565

Liu, Q., Gao, X., He, L., and Lu, W. (2017). Haze removal for a single visible remote sensing image. *Signal Process.* 137, 33–43. doi: 10.1016/j.sigpro.2017.01.036

Liu, S., Hu, Q., Li, P., Zhao, J., Liu, M., and Zhu, Z. (2019a). Speckle suppression based on weighted nuclear norm minimization and grey theory. *IEEE Trans. Geosci. Remote Sens.* 57, 2700–2708. doi: 10.1109/TGRS.2018.2876339

Liu, S., Ma, J., Yin, L., Li, H., Cong, S., Ma, X., et al. (2020). Multi-focus color image fusion algorithm based on super-resolution reconstruction and focused area detection. *IEEE Access* 8, 90760–90778. doi: 10.1109/ACCESS.2020.2993404

Liu, S., Wang, J., Lu, Y., Li, H., Zhao, J., and Zhu, Z. (2019b). Multi-focus image fusion based on adaptive dual-channel spiking cortical model in non-subsampled shearlet domain. *IEEE Access* 7, 56367–56388. doi: 10.1109/ACCESS.2019.2900376

Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., and Zhang, L. (2017). Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Trans. Image Process.* 26, 2519–2532. doi: 10.1109/TIP.2017.2671921

Mehrubeoglu, M., Smith, S., Simionescu, P. A., and McLauchlan, L. (2016). "Comparison of thermal and hyperspectral data to correlate heat maps with spectral profiles from galvanized steel surfaces," in *2016 IEEE/ACES International Conference on Wireless Information Technology and Systems (ICWITS) and Applied Computational Electromagnetics (ACES)* (Honolulu, HI), 1–2. doi: 10.1109/ROPACES.2016.7465386

Nejati, M., Karimi, M., Soroushmehr, S. R., Karimi, N., Samavi, S., and Najarian, K. (2017). "Fast exposure fusion using exposedness function," in *2017 IEEE International Conference on Image Processing (ICIP)* (Beijing), 2234–2238. doi: 10.1109/ICIP.2017.8296679

Nishino, K., Kratz, L., and Lombardi, S. (2012). Bayesian defogging. *Int. J. Comput. Vis.* 98, 263–278. doi: 10.1007/s11263-011-0508-1

Qi, G., Chang, L., Luo, Y., Chen, Y., Zhu, Z., and Wang, S. (2020). A precise multi-exposure image fusion method based on low-level features. *Sensors* 20:1597. doi: 10.3390/s20061597

Qin, X., Wang, Z., Bai, Y., Xie, X., and Jia, H. (2020). Ffa-net: Feature fusion attention network for single image dehazing. *Proc. AAAI Conf. Artif. Intell.* 34, 11908–11915. doi: 10.1609/aaai.v34i07.6865

Qing, C., Yu, F., Xu, X., Huang, W., and Jin, J. (2016). Underwater video dehazing based on spatial-temporal information fusion. *Multidimens. Syst. Sign. Process.* 27, 909–924. doi: 10.1007/s11045-016-0407-2

Rahman, Z.-u., Jobson, D. J., and Woodell, G. A. (2004). Retinex processing for automatic image enhancement. *J. Electron. Imaging* 13, 100–111. doi: 10.1117/1.1636183

Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* 38, 35–44. doi: 10.1023/B:VLSI.0000028532.53893.82

Rong, Z., and Jun, W. L. (2014). Improved wavelet transform algorithm for single image dehazing. *Optik* 125, 3064–3066. doi: 10.1016/j.ijleo.2013.12.077

Singh, D., and Kumar, V. (2017). Dehazing of remote sensing images using improved restoration model based dark channel prior. *Imaging Sci. J.* 65, 282–292. doi: 10.1080/13682199.2017.1329792

Tan, R. T. (2008). "Visibility in bad weather from a single image," in *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Anchorage, AK), 1–8. doi: 10.1109/CVPR.2008.4587643

Tarel, J.-P., and Hautiere, N. (2009). "Fast visibility restoration from a single color or gray level image," in *2009 IEEE 12th International Conference on Computer Vision (ICCV)* (Kyoto), 2201–2208. doi: 10.1109/ICCV.2009.5459251

Thomas, G., Flores-Tapia, D., and Pistorius, S. (2011). Histogram specification: a fast and flexible method to process digital images. *IEEE Trans. Instrum. Meas.* 60, 1565–1578. doi: 10.1109/TIM.2010.2089110

Wang, A., Wang, W., Liu, J., and Gu, N. (2019). AIPNET: Image-to-image single image dehazing with atmospheric illumination prior. *IEEE Trans. Image Process.* 28, 381–393. doi: 10.1109/TIP.2018.2868567

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Yu, L., Liu, X., and Liu, G. (2015). A new dehazing algorithm based on overlapped sub-block homomorphic filtering. *Eighth Int. Conf. Mach. Vis.* 9875:987502. doi: 10.1117/12.2228467

Yuan, X., Ju, M., Gu, Z., and Wang, S. (2017). An effective and robust single image dehazing method using the dark channel prior. *Information* 8:57. doi: 10.3390/info8020057

Zhang, H., and Patel, V. M. (2018). "Densely connected pyramid dehazing network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 3194–3203. doi: 10.1109/CVPR.2018.00337

Zheng, M., Qi, G., Zhu, Z., Li, Y., Wei, H., and Liu, Y. (2020). Image dehazing by an artificial image fusion method based on adaptive structure decomposition. *IEEE Sens. J.* 20, 8062–8072. doi: 10.1109/JSEN.2020.2981719

Zhu, Q., Mai, J., and Shao, L. (2015). A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.* 24, 3522–3533. doi: 10.1109/TIP.2015.2446191

Zhu, Z., Wei, H., Hu, G., Li, Y., Qi, G., and Mazur, N. (2021). A novel fast single image dehazing algorithm based on artificial multiexposure image fusion. *IEEE Trans. Instrum. Meas.* 70, 1–23. doi: 10.1109/TIM.2020.3024335

Zhu, Z., Zheng, M., Qi, G., Wang, D., and Xiang, Y. (2019). A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain. *IEEE Access* 7, 20811–20824. doi: 10.1109/ACCESS.2019.2898111

Conflict of Interest: QL, KL, WL, HD, and XJ was employed by company China Merchants Chongqing Communications Technology Research & 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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/fnbot.2021.739077

Research on a Segmentation Algorithm for the Tujia Brocade Images Based on Unsupervised Gaussian Mixture Clustering

*Shuqi He**

College of Computer Science, South-Central University for Nationalities, Wuhan, China

Tujia brocades are important carriers of Chinese Tujia national culture and art. It records the most detailed and real cultural history of Tujia nationality and is one of the National Intangible Cultural Heritage. Classic graphic elements are separated from Tujia brocade patterns to establish the Tujia brocade graphic element database, which is used for the protection and inheritance of traditional national culture. Tujia brocade dataset collected a total of more than 200 clear Tujia brocade patterns and was divided into seven categories, according to traditional meanings. The weave texture of a Tujia brocade is coarse, and the textural features of the background are obvious, so classical segmentation algorithms cannot achieve good segmentation effects. At the same time, deep learning technology cannot be used because there is no standard Tujia brocade dataset. Based on the above problems, this study proposes a method based on an unsupervised clustering algorithm for the segmentation of Tujia brocades. First, the cluster number *K* is calculated by fusing local binary patterns (LBP) and gray-level co-occurrence matrix (GLCM) characteristic values. Second, clustering and segmentation are conducted on each input Tujia brocade image by adopting a Gaussian mixture model (GMM) to obtain a preliminary segmentation image, wherein the image yielded after preliminary segmentation is rough. Then, a method based on voting optimization and dense conditional random field (DenseCRF) (CRF denotes conditional random filtering) is adopted to optimize the image after preliminary segmentation and obtain the image segmentation results. Finally, the desired graphic element contour is extracted through interactive cutting. The contributions of this study include: (1) a calculation method for the cluster number *K* wherein the experimental results show that the effect of the clustering number *K* chosen in this paper is ideal; (2) an optimization method for the noise points of Tujia brocade patterns based on voting, which can effectively eliminate isolated noise points from brocade patterns.

Keywords: Tujia brocade segmentation, GMM, DenseCRF, *K* auto-selection based on information fusion, optimization based on the vote

OPEN ACCESS

Edited by:
Shin-Jye Lee,
National Chiao Tung University, Taiwan

Reviewed by:
Amit Shakya,
Sant Longowal Institute of Engineering and Technology, India
Huan Deng,
Sichuan University, China
Jiasong Wu,
Southeast University, China

****Correspondence:***
Shuqi He
1229937232@qq.com

Received: *10 July 2021*
Accepted: *06 August 2021*
Published: *03 September 2021*

Citation:
He S (2021) Research on a Segmentation Algorithm for the Tujia Brocade Images Based on Unsupervised Gaussian Mixture Clustering.
Front. Neurorobot. 15:739077.
doi: 10.3389/fnbot.2021.739077

INTRODUCTION

Intangible cultural heritage is an important symbol of the historical and cultural achievements of a country or a nation. It is not only of great significance to the study of the evolution of human civilization but also plays a unique role in showing the diversity of world culture, being the common cultural wealth of mankind. Tujia nationality is one of the 56 ethnic groups in China. Tujia brocade is an important carrier of the culture and art of Tujia nationality. Furthermore, it records the most detailed and real cultural history of Tujia nationality, making it one of the National Intangible Cultural Heritage (Wan and Nie, 2018).

The basic primitives of a Tujia brocade are extracted by digital image technology for classification and storage to form a Tujia brocade database. This provides a safe and convenient way to protect the Tujia brocade culture. Tujia brocades use cotton yarn as warps and silk thread or cotton thread, and useful wool, as wefts, which are much thicker than ordinary fabric fibers. Therefore, the weave texture of a Tujia brocade is coarse, and it is not easy to form smooth and round curves and shapes. Brocade patterns have pixelated visual textures and the features of abstract geometric patterns (Wan and Nie, 2018). These characteristics make Tujia brocade images have exceptionally large color characteristic differences from ordinary images, and the texture-level image contrast is not strong, which brings difficulty to image segmentation.

Image segmentation is one of the research hotspots in the field of computer vision. The traditional image segmentation algorithms mainly use the low-level semantics of images including color, texture, and shape for segmentation, such as threshold method, region grow algorithm, and edge detection algorithm, among others (Heath et al., 1997; Fan et al., 2001; Otsu, 2007). Superpixel segmentation methods emerged after researchers introduced graph theory to image segmentation such as Graph Cuts and Simple Linear Iterative Clustering (SLIC) (Felzenszwalb and Huttenlocher, 2004; Achanta et al., 2012). It is difficult to achieve semantic segmentation *via* traditional clustering segmentation based on the shallow features of images.

The model based on deep learning can automatically extract the image features representation and has achieved excellent results in many challenging computer vision tasks, including object detection, location, recognition, and segmentation. Classic image segmentation models such as Fully Convolutional Networks (FCN) (Long et al., 2015), Mask Regional-Based Convolutional Neural Networks (Mask R-CNN) (He et al., 2017), DeepLab, and so on. The semantic segmentation DeepLab (Chen et al., 2018a,b) employs a series algorithm by integrating various classical deep learning methods and using Atrous Convolution, Atrous Spatial Pyramid Pooling (ASPP), along with the other structures. Meanwhile, a dense conditional random field (DenseCRF) structure was connected to the back end of the neural network to provide a refined segmentation for the boundary after initial segmentation. Nonetheless, most classic image segmentation models rely on high-quality massive datasets. It is difficult to conduct image segmentation by the classic deep learning segmentation model because the dataset in this study only contains more than 200 images without a pixel-level segmentation tag.

More recently, unsupervised deep learning becomes a research hotspot. A dual-branch combination network (DCN) (Yang et al., 2017) was proposed as a method combining an autoencoder and K-means. The model encoder maps the input data from high-dimensional features to low-dimensional subspaces, obtains the potential features of the data, performs K-means clustering on them, and obtains the K-means loss. The decoder reconstructs the latent features into the original data to obtain the reconstruction loss. The network combines the reconstruction loss and K-means loss through backpropagation to optimize the learning process. The study of Kanezaki (2018) used standard unsupervised over-segmentation techniques to supervise convolutional neural networks. This method uses standard algorithms to extract pre-segmented regions from the original image. The segmentation model extracts image features through convolutional neural networks to obtain a rough segmentation of the image and then adjusts the rough segmentation results according to multiple constraints, such as feature similarity and spatial continuity so that all pixels in the same pre-segmented area have the same label. The loss incurred between the two segmentation images before and after the adjustment is used as the backpropagation loss of the supervision signal to update the network weight.

The recognition and segmentation of brocade texture are also one of the applications of image segmentation. Brocade texture feature extraction technology originated in the mid-1980's. Over the past decade, researchers began to focus on textile-aided design, fabric pattern segmentation, and contour extraction technology. The study of Kuo et al. (2005, 2007) and Kuo and Shih (2011) advocated extracting the color features of printed fabrics through feature extraction algorithms, such as self-organizing map network (SOM), and then obtained the pattern by using the Fuzzy-c means (FCM) algorithm to achieve the automatic classification of the colors. The study of Lachkar et al. (2006) adopted a clustering method based on a GMM. The method combined a GMM and a content validity index (CVI) to form an adaptive, efficient segmentation algorithm. In the research conducted by Jiang et al. (2014), they studied the automatic recognition technology of jacquard warp knitted fabric pattern images. The fabric image uses a two-dimensional wavelet decomposition algorithm to extract features, given the clustering center, and then uses the K-means clustering multi-channel algorithm for segmentation.

Based on the research of textile image segmentation algorithm, we found that there are two difficulties in the segmentation of Tujia brocade by the commonly used image segmentation algorithm.

- The material of Tujia brocade is rougher than the common fabric fiber and the background texture of the brocade pattern is very prominent. This forms a similar feeling to "mosaic," which is represented as a noise signal on the fabric image. Such kind of noise information can cover up part of the detail information, and increase the image entropy, making the boundary between the Tujia brocade primitive and the image background becomes blurred. This will increase the difficulty of edge detail segmentation and reduce the accuracy of pattern texture segmentation.

- Deep learning-based image segmentation algorithms typically use large datasets for training to prevent overfitting during data processing. Tujia brocade image segmentation research is relatively rare; there is a lack of training data specifically designed for brocade image segmentation. If the image matting or image segmentation tools are used to build a data set, it needs a lot of manual labor to extract material from massive data through tedious operations.

In response to the above problems, this study proposes a clustering segmentation process for Tujia brocades. First, the input Tujia brocade is divided into basic clusters. Then, a voting-based optimization method is used to eliminate the noise points of the image based on the characteristics of the Tujia brocade. Afterward, DenseCRF is employed to optimize the image and obtain effective segmentation results. Finally, the desired primitive outline is extracted through interactive cutting. The contributions of this study are as follows:

- A calculation method for cluster number *K*. In unsupervised clustering, the *K*-value has an extraordinarily strong impact on the clustering results. The algorithm uses local binary patterns (LBP) to calculate the base for the image texture features and uses the feature value of the GLCM as the weight. The two values are fused to calculate the *K*-value for clustering. Experiments show that the clustering effect of the *K*-value selection algorithm is ideal.
- An optimization method for Tujia brocade noise points. Due to the extensive weave textures of Tujia brocades and the obvious textural characteristics of the background, noise points easily occur after clustering. DenseCRF can be used to optimize the image contour, but it is not effective in eliminating the noise points of a Tujia brocade. Therefore, we propose a voting-based optimization method. The classification labels obtained after the preliminary clustering process are voted on according to the classification results of their neighboring pixels to redistribute the labels of the center pixels. This method for the elimination of isolated noise points is remarkably effective and is then combined with DenseCRF to optimize the preliminary clustering-based segmentation map to obtain the final Tujia brocade segmentation map.

METHOD

For a small unlabelled dataset, we used an unsupervised clustering method to segment the input Tujia brocade. First, the LBP and GLCM feature values were fused to calculate the *K*-value of the cluster. Afterward, a GMM is used to cluster and obtain a preliminary segmentation map. This approach does not extract image features that are different from those obtained *via* traditional image segmentation. The image yielded after the initial segmentation process is relatively rough, and we propose a method based on the combination of voting optimization and DenseCRF to optimize this to obtain the final image segmentation result. The specific flow chart is shown in **Figure 1**.

Cluster Number *K* Auto-Selection

In an image, regions belonging to the same object mostly have similar textures and colors. During the image clustering segmentation, similar pixels were classified into a category. This category is regarded as a segmentation object which is classified according to the similarity between image pixels. The *K*-value selection is particularly important to obtain a good image segmentation effect. Due to the influence of brocade weaving technology, the image background of Tujia brocade has a strong sense of grain. If the *K*-value is too large when clustering, the image background will be clustered, forming the mosaic effect and affecting the segmentation effect. However, if the *K*-value is too small, the fine lines in the image will be ignored. **Figure 2** shows the segmentation effect of different *K*-values in the GMM algorithm.

Under observation, we found that the visual effect of the clustering was better when $K = 2$, 3, or 6, but we were not sure exactly what the clustering *K*-value should be until the clustering results come out. The model was selected mostly through criterion functions such as Bayesian information criterion (BIC) (Chakrabarti and Ghosh, 2011), Akaike information criterion (AIC) (Burnham and Anderson, 2002), among others. However, such application was very difficult in the actual model selection because the computational effort was too large, and it was found *via* specific experiments that the model selected by the criteria function was not the optimal estimation model for the image segmentation. All models obtained by training were only regarded as an approximate model of the real model. The objective of this study is to obtain a reasonable clustering *K*-value quickly and effectively. Traditional Tujia brocade consists of many similar graphic elements with strong regularity and has obvious texture features. For this reason, the number of texture features can be used to select the *K*-value of the clustering model. We introduce the statistical eigenvalues of the image GLCM and LBP to calculate the *K*-value.

Local Binary Patterns

Local binary patterns is an operator to describe the local texture features of the image and has gray and rotation invariance. LBP operator proposed by the study of Ojala et al. (2002) can divide the whole image into different subregions to perform local texture feature histogram statistics in each small region, that is, to count the feature number belonging to a certain pattern in the region. Finally, the histogram of all regions was connected as the image feature vector. The original LBP operator took the center pixel of the 3 × 3 window as the threshold value to compare the gray values of the adjacent eight pixels with the threshold value in turn clockwise. If the gray value is greater than or equal to the threshold value, the value of this pixel point is marked as 1, otherwise 0. After the comparison between the adjacent eight pixels, an 8-bit binary number was generated as the LBP value of the center pixel of the window to reflect the texture information

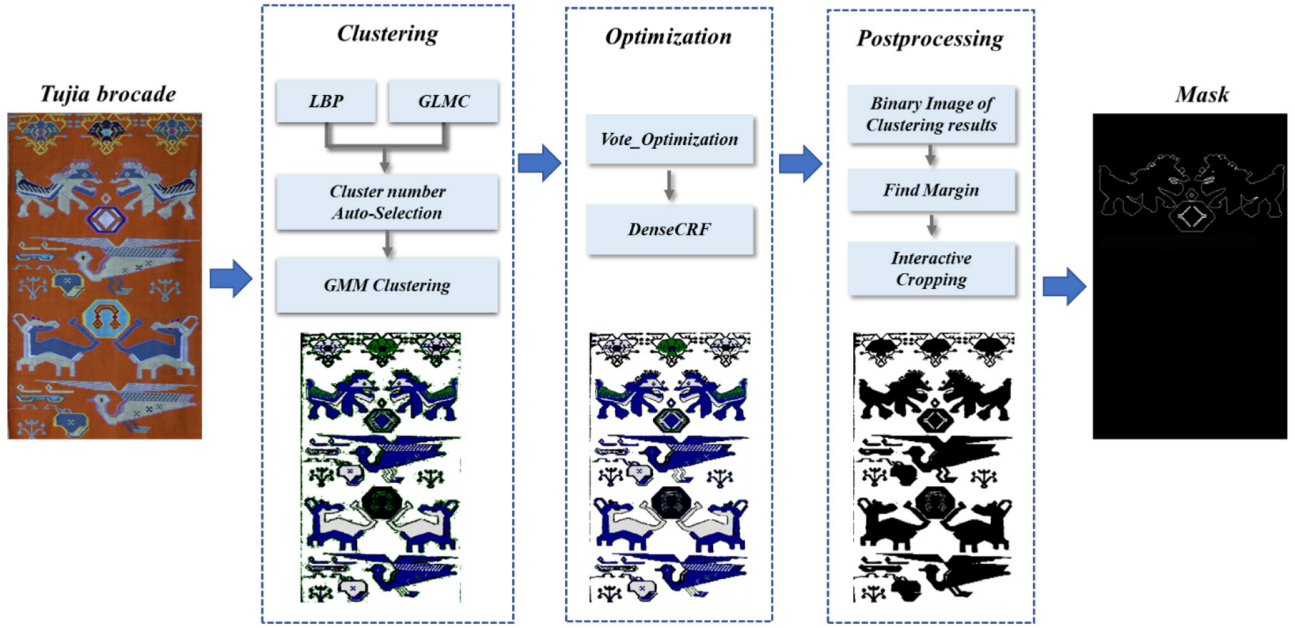


FIGURE 1 | Tujia brocade image segmentation process.

of the region. The specific calculation process is shown in Formula (1).

$$LBP\left(x_c, y_c\right) = \sum_{p=0}^{p-1} 2^p s\left(i_p - i_c\right). \tag{1}$$

where (x_c, y_c) is the coordinate of the central pixel; p is the p^{th} pixel of the adjacent region; i_p is the gray value of the pixel of the adjacent region; i_c is the gray value of the central pixel; *s(x)* is a sign function as shown in Formula (2).

$$S\left(x\right) = \begin{cases} 1, & if \;\; x \geq 0 \\ 0, & \quad else \end{cases} \tag{2}$$

The original LBP operator only covers a small area of 3 × 3 in practical application, which cannot adapt to the texture features of different sizes. For this purpose, Extended LBP (Ojala et al., 2002) was proposed which extended the coverage area of the LBP operator to a circular neighborhood with a radius of *R*. The LBP operator can sample P points in the circular region. The method adopted Uniform Pattern LBP. P sampling points generated 2^P patterns in Extended LBP. The introduction of "equivalent mode" (Ojala et al., 2002) reduced the number of modes from the original 2^P to P(P – 1) + 2. We adopted the LBP algorithm which can calculate the occurrence frequency of image texture feature pattern, to calculate the cardinality of clustering *K*-value.

Gray Level Co-occurrence Matrix (GLCM)

Tujia brocade images are generally permuted by many repeated arrays of basic primitives. The basic texture feature cardinality calculated by the LBP operator may not fully represent the number of categories of segmented objects. Therefore, we introduced the statistical feature values of the image GLCM (Sulochana and Vidhya, 2013) which was commonly used to describe texture by studying the spatial correlation characteristics of gray level. The texture is formed by the repeated appearance of gray distribution in spatial positions, so there is a certain gray relationship between two pixels separated by a certain distance in the image space, that is, the spatial correlation characteristics of gray level in the image. For GLCM, the joint probability density of the two pixels was used to reflect the gray direction, interval, and change amplitude of the image. However, GLCM cannot directly provide the features of the texture. Some scalars can be used to represent GLCM features. The entropy value of the co-occurrence matrix contains the randomness measure of the image information amount, indicating the complexity of the image gray level distribution. The greater the entropy value is, the more complex the image is, as shown in the calculation Formula (3). The *M*-value reflects the degree of regularity of the texture. The smaller *M*-value means that the texture features are more chaotic and difficult to describe, as shown in the calculation Formula (4). The greater the contrast of the image, the clearer the visual effect of the image, as shown in the calculation Formula (5). We assumed that images with more complex patterns and chaotic texture features tended to be described by more models.

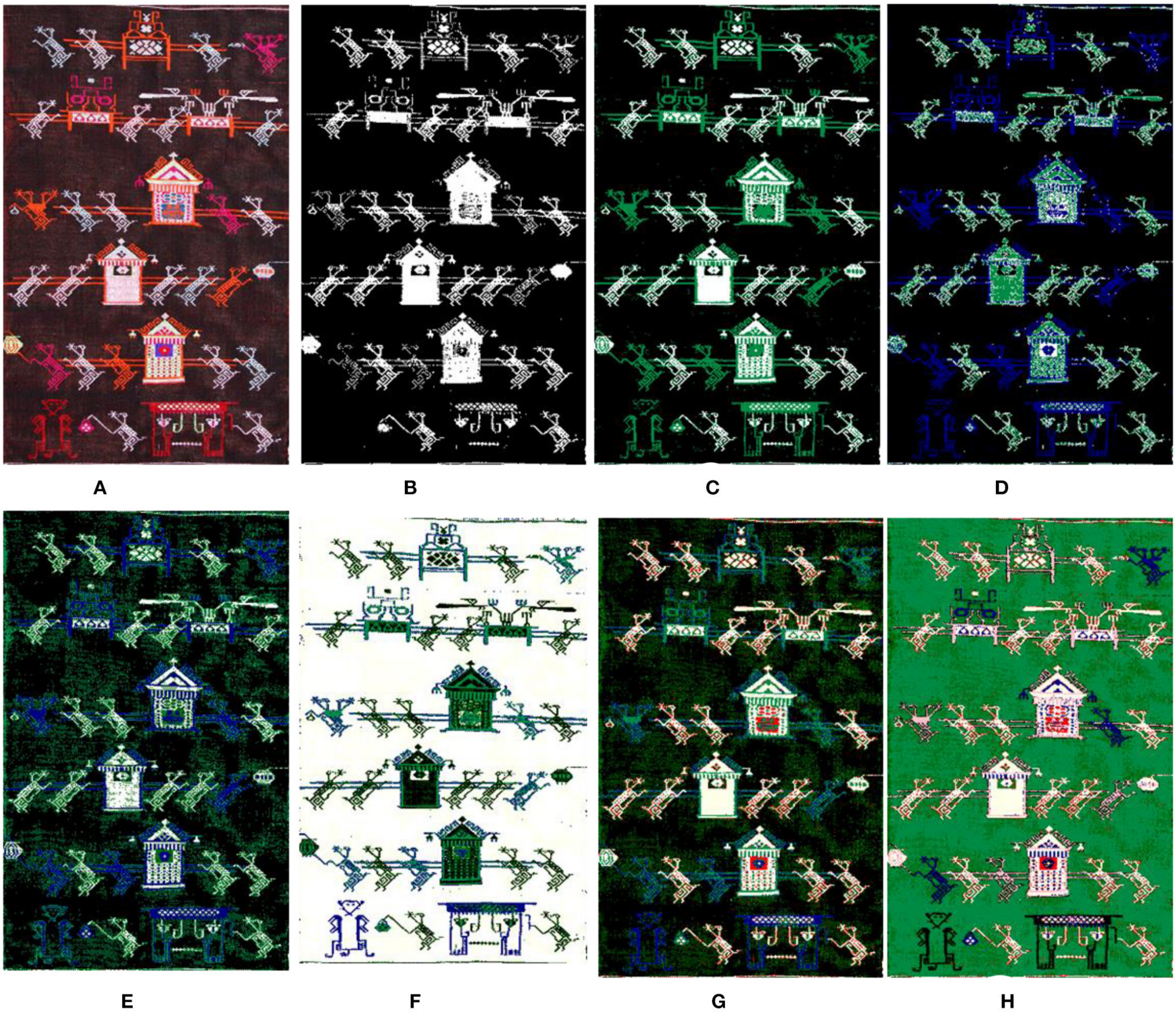


FIGURE 2 | The segmentation results by the GMM uses different K-values. **(A)** Original. **(B)** $K = 2$. **(C)** $K = 3$. **(D)** $K = 4$. **(E)** $K = 5$. **(F)** $K = 6$. **(G)** $K = 7$. **(H)** $K = 8$.

$$Entropy = -\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P\left(i,j,d,\theta\right) \times \ln P\left(i,j,d,\theta\right) \tag{3}$$

$$Mean = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P\left(i,j,d,\theta\right) \times i \tag{4}$$

$$Contrast = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p(i,j) \times (i-j)^2 \tag{5}$$

Calculating K-Values

The occurrence frequency of LBP texture features in the image was counted by the algorithm where a threshold value was set up and the number of LBP features whose frequency exceeds the threshold value was used as the cardinality of clustering K-value. Entropy, M, and contrast parameters of GLCM were used to calculate the weight of the clustering K-value. The calculation formula of K-value was shown as Formula (6). The weight calculation formula of clustering K-value was shown as Formula (7).

$$K = COUNT\left(P(LBP_{image\_i}) > threshold\right) \times W_{image\_i} \tag{6}$$

$$W_{image\_i} = Entropy \times \alpha_1 + Mean \times \alpha_2 + Contrast \times \alpha_3 \tag{7}$$

In Formula (6), $P(LBP_{image\_i})$ represents the frequency of a texture feature; W_{image_i} represents the image texture complexity measure of *image_i*, which is obtained by Formula (7).

Gaussian Mixture Model

The GMM (Bishop, 2006) is a probabilistic model. In image segmentation, image features, such as gray information, color information, or texture information, are used as the observation vectors of the image. It is assumed that the overall image pixels obey a Gaussian mixture distribution. The segmented areas can be regarded as single Gaussian models with the same form, and each model is independent of all other models. The entire image is a GMM formed by fusing multiple single Gaussian models with a certain weight.

Assuming that the GMM is composed of K-Gaussian models (the data contain K-classes), the probability density function of the GMM is shown in Formula (8) (Bishop, 2006).

$$p\left(x\right)=\sum_{k=1}^{K}w_k g(x|\mu_k,\sum\nolimits_k) \tag{8}$$

where K is the number of components in the GMM, w_k is the mixture weight, which represents the proportion of the K single Gaussian models in the mixture model, $0\leq w_k\leq 1$, $\sum_{k=1}^{K}w_k=1$, $g(x|\mu_k,\sum_k)$ is the distribution of the Gaussian component k, and its function expression is shown in Formula (9) (Bishop, 2006).

$$g(X)=\frac{1}{\sqrt{(2\pi)^N|\sum|}}e^{-\frac{1}{2}(X-\mu)^T\sum^{-1}(X-\mu)} \tag{9}$$

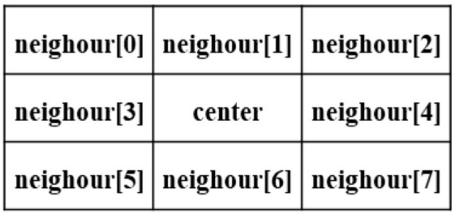
where X is a random variable (which can be understood as the observation vector of the image), N is an arbitrary integer determined by the dimensionality of X, μ is the mean vector, $\mu=E\{X\}=[\mu_1,\mu_2,\cdots,\mu_N]^T$, $\sum$ is the covariance matrix, N × N represents the number of dimensions, $\sum^{-1}$ is the inverse matrix of $\sum$, and $|\sum|$ is the determinant of $\sum$, $\sum=E\left\{(X-\mu)(X-\mu)^T\right\}$. The iterative EM algorithm is used to solve the likelihood function criterion of the GMM and estimate the Gaussian distribution parameters to obtain the probability that each pixel belongs to each category. Finally, the category with the highest probability is regarded as the category to which the pixel belongs; this process is repeated until all image pixels have been classified, thus realizing the segmentation of the entire image. The likelihood function criterion is shown in Formula (10) (Bishop, 2006).

$$L\left(\theta\right)=ln\left[\prod_{i=1}^{n}p\left(x\right)\right]=\sum_{i=1}^{n}ln\sum_{k=1}^{K}w_j g(x|\mu_k,\sum\nolimits_k) \tag{10}$$

Optimization of the Clustering Results

Optimization Method Based on the Voting Method

Traditional Tujia brocades use cotton yarn, silk thread, or cotton thread as the main weaving materials, and the formed image background has a strong weave texture, as shown in **Figure 4A**. After clustering, some noise points are formed that affect the segmentation results, as shown in **Figure 4B**. GMM clustering yields the classification probabilities of image pixels. Clustering does not consider the relationships between image pixels, and misclassification occurs when the image quality is not high. Generally, adjacent pixels in an image may belong to the same object. We draw on the idea of voting and define a 3 × 3 window, as shown in **Figure 3**. The center pixel is reassigned to a category according to the classification probabilities of the adjacent eight pixels. The algorithm sets a threshold, and when the probability of a category among the eight pixels adjacent to the center pixel exceeds the threshold, the category of the center pixel is modified to this category.



neighour[0]	neighour[1]	neighour[2]
neighour[3]	center	neighour[4]
neighour[5]	neighour[6]	neighour[7]

FIGURE 3 | Optimization based on a voting window.

Each *neighbor[i]* has two attributes (Prob, label), where the label represents the assigned category *k* for the pixel, *Prob=[P(2), P(3), ..., P(k)]*, *k* ∈ *[2, K]*, and *P(k)* represents the probability that this pixel belongs to category *k*. The category assignment of the center pixel is calculated *via* Formula (11).

$$Prob[k]=\max_{k\in(1,K)}\left(Average\sum_{i=1}^{8}Prob\left(neighour\left[i\right]\right)\right) \tag{11}$$

If *Prob[k]* > = *threshold*, then *label[center]* = *k*.

Experiments show that this optimization process has a good effect on eliminating obvious independent noise points. As shown in **Figure 4C**, after the algorithm iteratively optimizes the image once, the background lines and noise points evidently disappear.

Dense Conditional Random Field

Optimization based on the voting method considers only the associations between neighboring pixels without considering the overall image and cannot optimize the image globally. As shown in **Figure 4C**, the details of the clustering result are relatively rough. For further optimization, we introduce DenseCRF. If the distance between and colors of the image pixels are very close, they belong to the same category in theory. DenseCRF (Philipp and Koltun, 2012) readjusts the existing clustering results from these two aspects based on the colors and the spatial location information of the pixels provided by the entire image and assigns the attributes of the pixels. In the fully-connected random field, the energy function of label x is expressed as Formula (12) (Philipp and Koltun, 2012).

$$E\left(x\right)=\sum_{i}\theta_i\left(x_i\right)+\sum_{ij}\theta_{ij}\left(x_i,x_j\right) \tag{12}$$

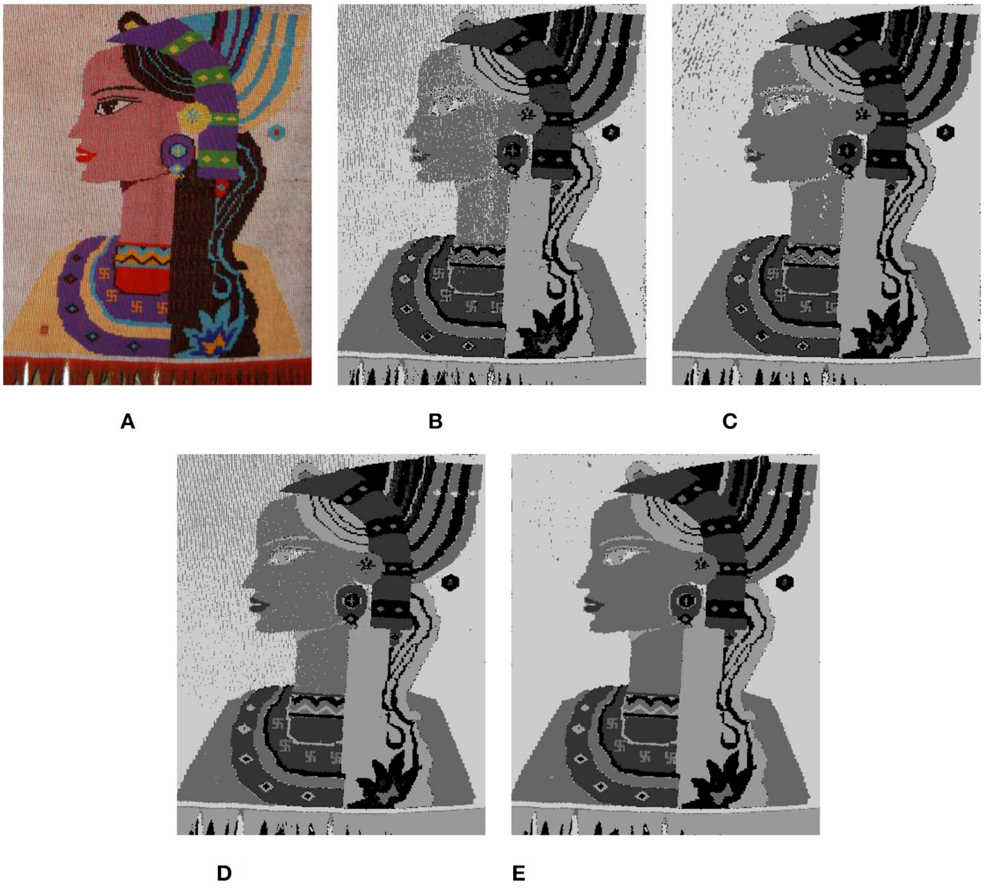


FIGURE 4 | Optimization of clustering results. **(A)** Image: Original. **(B)** GMM, $K = 5$. **(C)** Vote_Optimization: GMM + Vote_Optimization. **(D)** GMM + DenseCRF. **(E)** GMM + DenseCRF + Vote_Optimization.

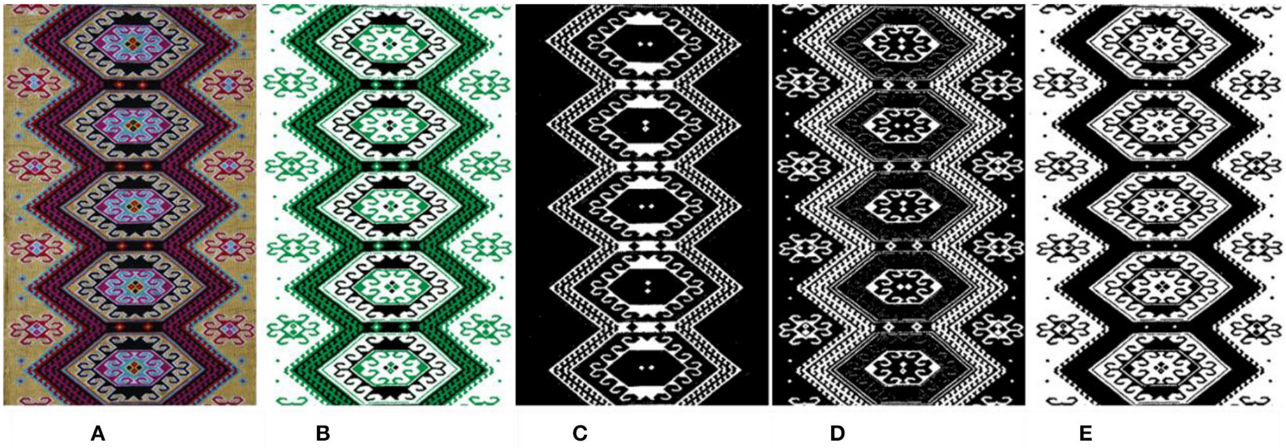


FIGURE 5 | Binarized images of the clustering categories. **(A)** Image: Original. **(B)** Cluster ($k = 3$). **(C)** Binary Image (Label = 0). **(D)** Binary Image (Label = 1). **(E)** Binary Image (Label = 2).

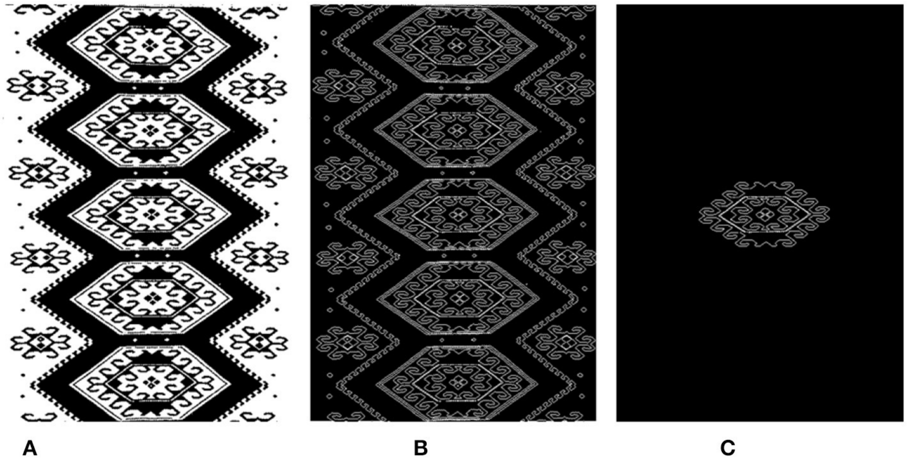


FIGURE 6 | Image mask obtained by interactive cutting. **(A)** Binary Image (Label = 2). **(B)** Contour. **(C)** Mask.

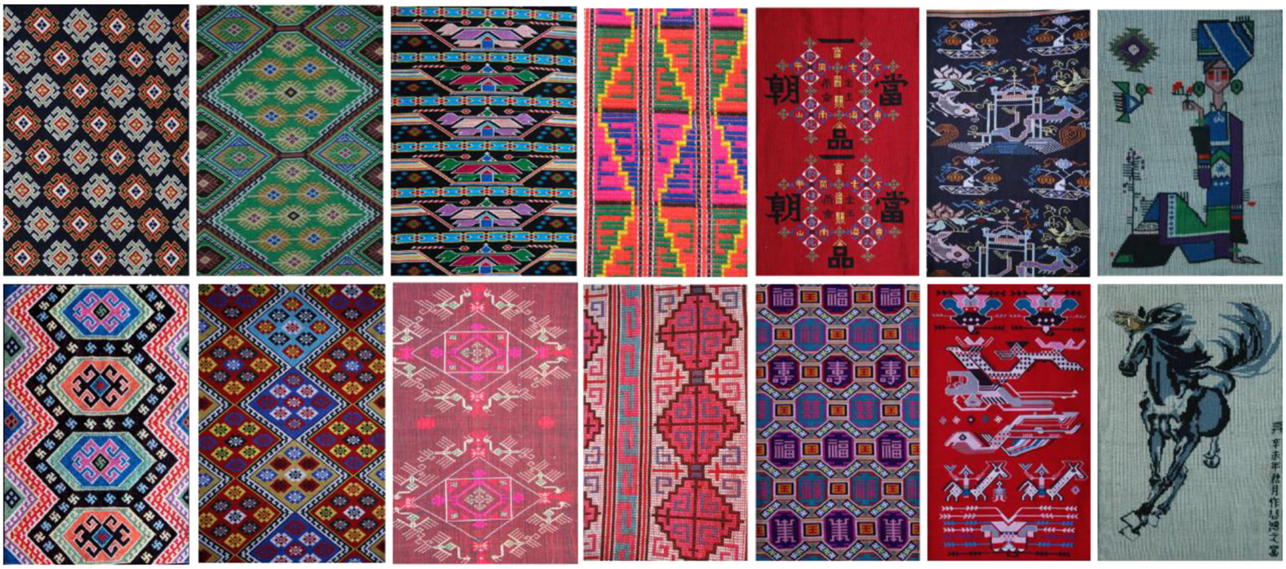


FIGURE 7 | Tujia brocade dataset.

In the formula, the unary potential $\theta_i(x_i)$ comes from the front-end output (such as predicted by a classifier), and it represents the energy of dividing pixel i into label x_i, which includes the shape, texture, position, and color of the image. The pairwise potentials $\theta_{ij}(x_i,\ x_j)$ is the energy in which the pixel i and j are simultaneously assigned label x_i and x_j. It describes the relationship between the pixels and encourages similar pixels to be assigned the same label. Pixels with large differences are assigned different labels so that the model can segment the image at the boundary as much as possible.

As shown in **Figure 4D**, the details of the clustering results are more delicate and smoother after DenseCRF optimizes the clustering results, but there are still background textures and noise points. We combine the two optimization methods, and the final optimization result is shown in **Figure 4E**.

Mask Extraction

A Tujia brocade is a geometric lattice pattern. Because of the interweaving of warps and wefts, its patterns are mostly composed of parallel lines, vertical lines, and diagonal lines. The clustering results in **Figure 5B** are shown in **Figures 5C–E**, which correspond to the binarized images (black background) of the clustering categories (such as label = 0 and label = 1). Each binary image (as **Figure 6A**) can be regarded as a part of the texture object that needs to be extracted, and its contour (as

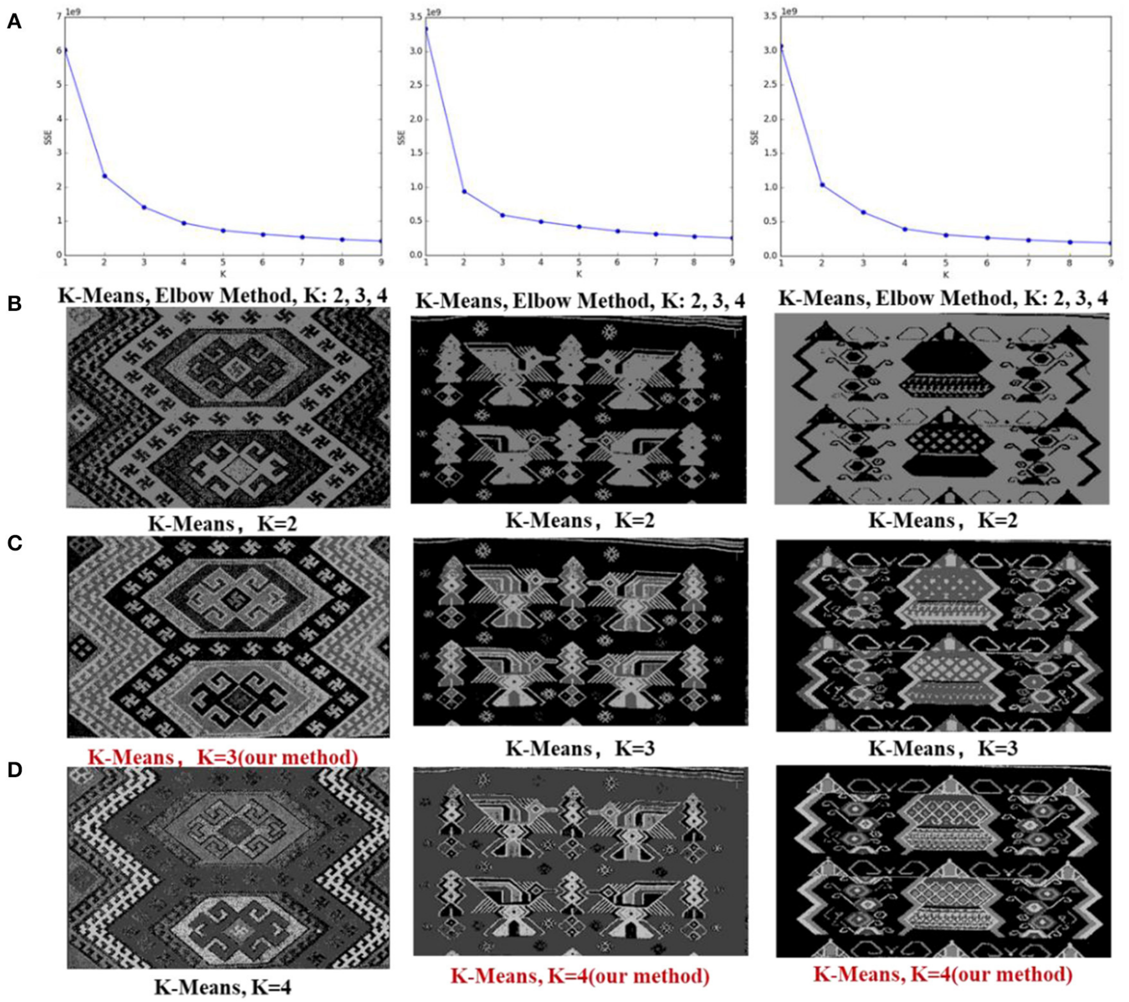


FIGURE 8 | (A) Elbow method curve diagram; **(B–D)** *K*-means clustering results based on the *K*-values obtained by the elbow method. The red color represents the *K*-values calculated by the algorithm in this study.

TABLE 1 | Calinski-Harabaz index (CH) ranking table.

CH ranking	1	2	3	4	5	6	7	8
Our method (sheet)	61	12	14	10	14	9	16	20

TABLE 2 | The calculation time of the cluster value *K*.

Algorithm	Elbow method	Calinski-Harabaz	Our method
Time to calculate *K*-value (sheet)	40.17 s	56.05 s	0.22 s

Figure 6B) is detected for interactive segmentation to obtain the object mask, as shown in **Figure 6C**.

EXPERIMENTS

The experiment is implemented by Python Software Foundation and the experimental environment is Microsoft Windows 10. The testing machine contains an Intel Core i7-8750H 2.20 GHz, an Nvidia GeForce GTX 1060 with Max-Q Design, and 24 GB of memory.

Dataset

Since there are few studies on Tujia brocade image segmentation based on machine learning, there is no ready-made Tujia

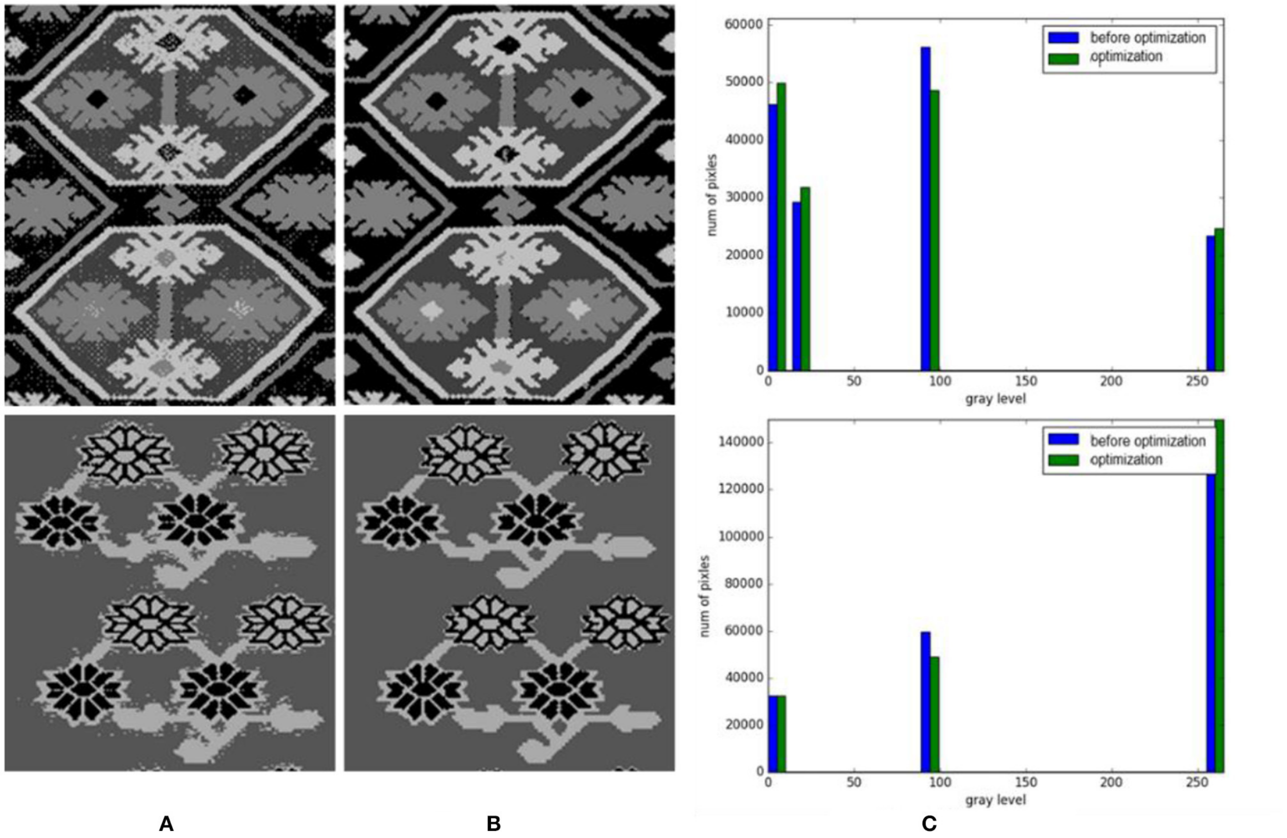


FIGURE 9 | Gray histograms of cluster contrast optimization. **(A)** Before optimization. **(B)** Optimization. **(C)** Gray histograms.

brocade dataset for use in experiments. We retrieve public Tujia brocade image data from the Internet, manually photograph the Tujia brocade, and collect a total of more than 200 clear Tujia brocade patterns. According to the traditional meanings of the Tujia brocades, the patterns are roughly divided into six categories: animals patterns, flowers and plants patterns, living utensils patterns, natural object patterns, geometric patterns, text patterns. The woven material of a Tujia brocade is rougher than ordinary fabric fibers, so the background textures of the brocade patterns are very prominent, the pictures are not clear, and the brocades are bright in color, as shown in **Figure 7**.

Selection of the Clustering Value *K* and Evaluation of the Clustering Results

Evaluation by the Elbow Method

In unsupervised clustering, the clustering effect on the image details becomes clearer as the K-value increases, which is due to the particularity of the Tujia brocade dataset. When K reaches a certain critical point, the definition of the image details increases. However, the background texture is also clustered, forming noise points that affect the clustering results.

In the experiment, the cluster value K is calculated by auto-selection. To verify whether the selection of the K-value produced by the algorithm is reasonable, the K-means algorithm is used to conduct an experimental comparison on 100 Tujia brocade pictures. Based on the index of the intra-cluster error variance [the sum of squared errors (SSE)] through the elbow method (Marutho et al., 2018), different K-values ($K \in [2,9]$) are selected to repeatedly train multiple K-means models to obtain relatively suitable clustering categories. The output values are then compared with the K-values calculated by the algorithm. **Figure 8A** displays the clustering SSE line graph obtained by the elbow method algorithm. As shown in **Figure 8A**, the optimal range of k-value is 2,3,4. **Figures 8B–D** shows the segmentation results of $k = 2,3,4$.

Calinski-Harabaz Index (CH)

For a clustering task, because the structure of the given dataset is unknown, the evaluation of the clustering results relies only on the characteristics and values of the dataset itself. Usually, the density within each cluster and the degrees of dispersion between clusters are used to evaluate the effect of clustering. Commonly used evaluation indicators are the silhouette coefficient (Luan et al., 2012) and CH (Liu et al., 2020). The CH is simple to calculate and runs much faster than the silhouette coefficient. Therefore, we choose the CH to evaluate the clustering effect of

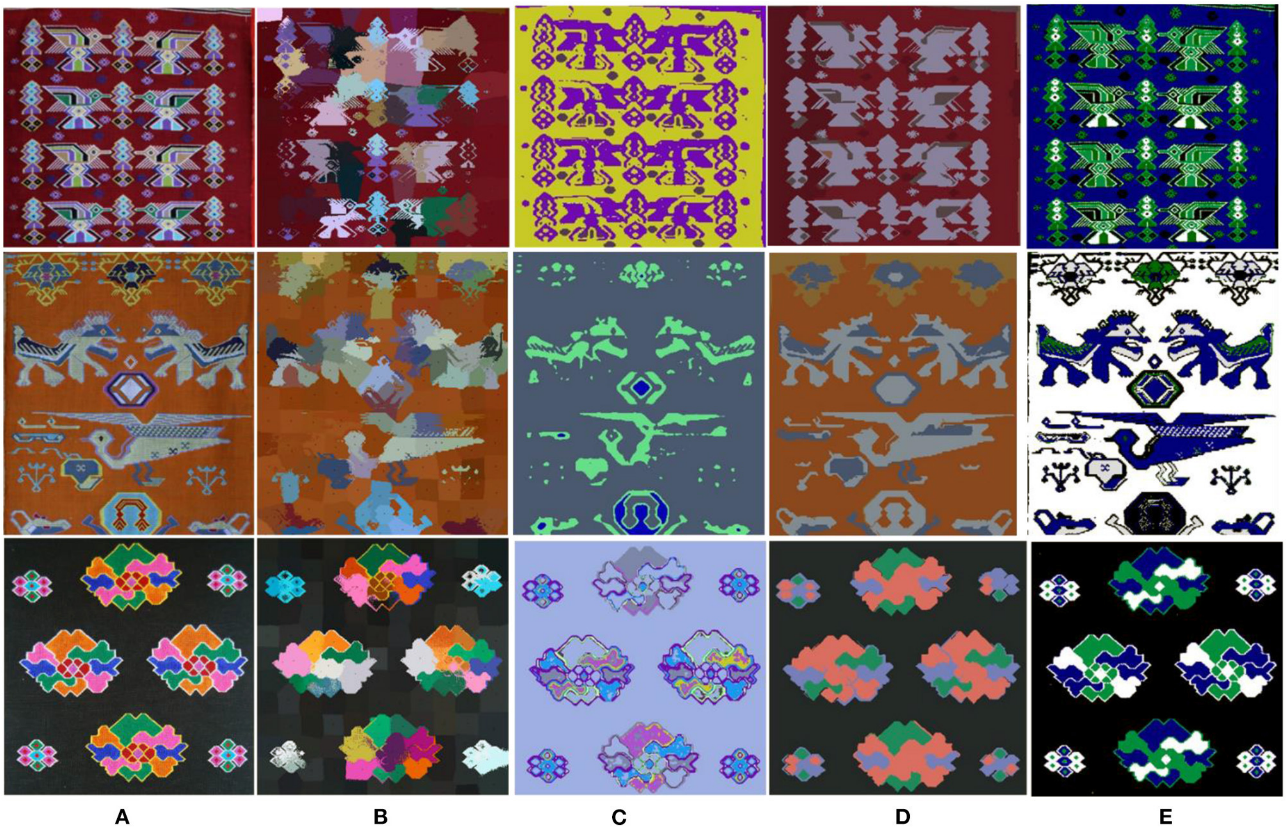


FIGURE 10 | Contrast classic image segmentation algorithms and the method. **(A)** Original. **(B)** SLIC. **(C)** DCN. **(D)** Unsupervised Image Segmentation. **(E)** Our method.

the approach. The CH calculation formula (13) (Liu et al., 2020) is as follows.

$$CH\left(k\right)=\frac{trB\left(k\right)/\left(k-1\right)}{trW\left(k\right)/\left(n-k\right)} \tag{13}$$

where n represents the number of clusters, k represents the current class, *trB(k)* represents the trace of the inter-class dispersion matrix, and *trW(k)* represents the trace of the intra-class dispersion matrix. The larger the CH, the tighter the class itself, and the more scattered the classes, better clustering results are obtained.

In the experiment, the CH is calculated based on 156 Tujia brocades, and GMM is used to calculate the CH value of each cluster from $k = 2$ to $k = 9$. The K-value rankings of our method are shown in **Table 1**.

The commonly used clustering value selection methods and the method in this article are compared in terms of their running times and are shown in **Table 2**.

Among them, the CH score for the K-value calculated by the method is the highest at 61. However, the highest CH score does not necessarily correspond to the best visual effect due to the particularity of the Tujia brocade dataset.

Cluster Segmentation and Optimization Results

It was found through the experiments that K-means clustering is extremely sensitive to the choice of the K-value; K-means is also sensitive to noise points. The clustering effect is very good when the image background is clear and monotonous, but the clustering effect is not very good if the optimal cluster value K is not chosen or the image background texture is not obvious. Comparing the experimental results, it is found that GMM is more robust to the dataset than other models. As long as a suitable K-value range is chosen, the clustering effect is improved and the background texture characteristics have relatively little effect on the clustering results. From the perspective of the entire dataset, the GMM clustering effect is better than the K-means effect on the whole dataset.

Due to the particularity of Tujia brocade material and the brocade process, some noise points are formed after image clustering that affects the segmentation results. Therefore, we optimize the results after image clustering and compare the greyscale histograms before and after image optimization. The results are shown in **Figure 9**.

We adopted some classic image segmentation algorithms, such as SLIC (Achanta et al., 2012), DCN (Yang et al., 2017),

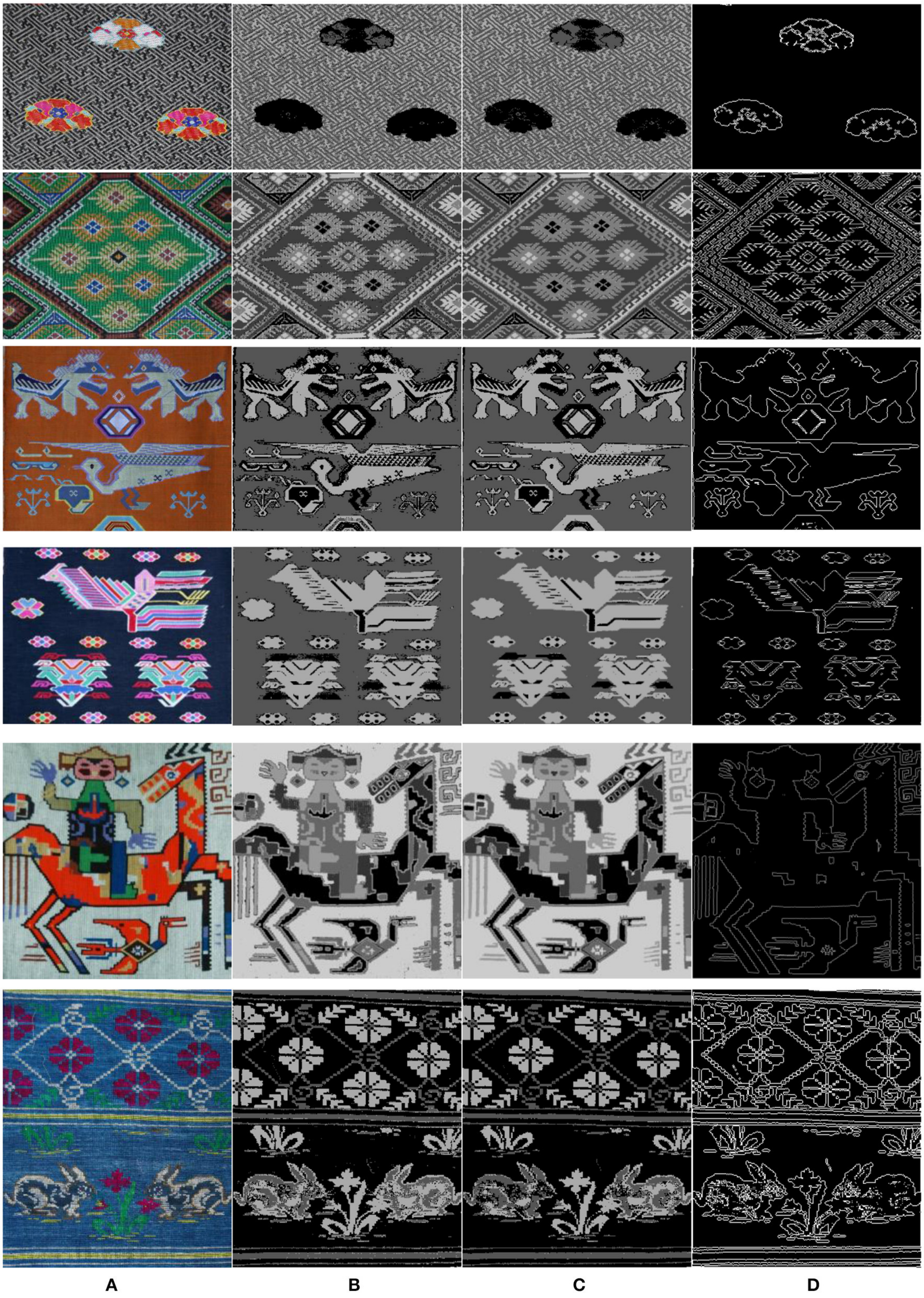


FIGURE 11 | Clustering and optimization results of Tujia brocades. **(A)** Image: Original. **(B)** Cluster. **(C)** Optimization. **(D)** Mask.

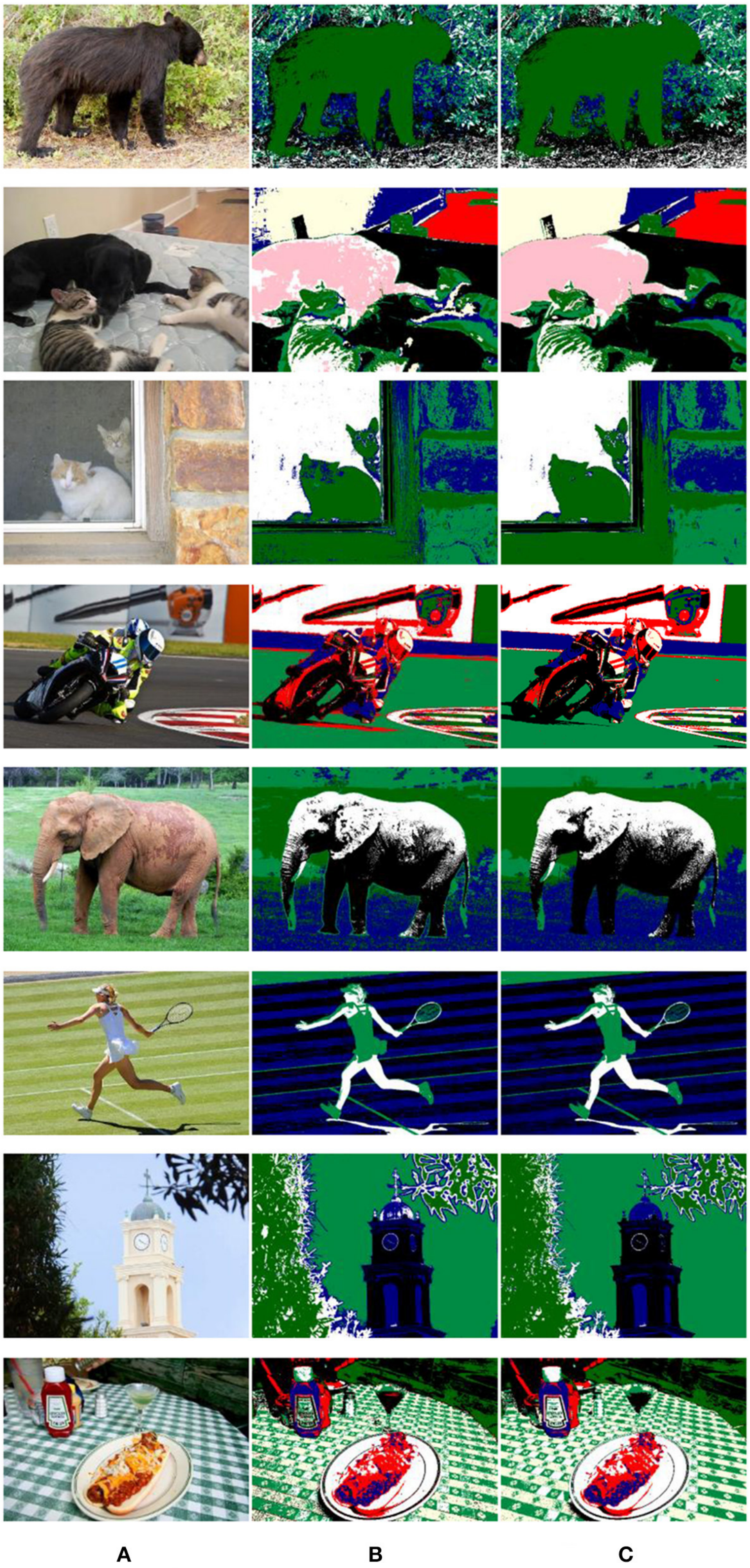


FIGURE 12 | Clustering and optimization results of natural scene pictures. **(A)** Image: Original. **(B)** Cluster. **(C)** Optimization.

and Unsupervised Image Segmentation (Kanezaki, 2018), and the algorithm proposed in this study for the image segmentation of Tujia brocade. The segmentation results are shown in **Figure 10**. It was revealed that image information was lost by the segmentation based on a convolutional neural network as shown in **Figures 10C,D**.

Figure 11A is the original picture, **Figures 11B,C** show the clustering and optimization results of some Tujia brocades. After the clustering and optimization processes are completed, the required object mask is extracted, and the specific result is shown in **Figure 11D**.

Figure 12A is a randomly selected picture from the Microsoft Common Objects in Context (MS COCO) dataset. The K-value of the cluster is calculated by the algorithm proposed in this article, and then the GMM is used for clustering. The result is shown in **Figure 12B**. The images in the MS COCO dataset are all high-definition pictures, and there is less interference from noise points, so only the dense conditional random field (DenseCRF) method is used in the optimization process, and the result is shown in **Figure 12C**.

CONCLUSION

Due to the lack of a segmentation dataset for Tujia brocades, this article uses an unsupervised clustering method to segment Tujia brocades. Due to the rough textures of Tujia brocade patterns, the clustering results are more sensitive to the K-value, so we propose a K-value auto-selection algorithm based on a GLCM and LBPs. This method can quickly and effectively calculate a suitable K-value, and the speed is close to 100 times that of the elbow method and the CH approach. At the same time, an optimization method based on voting is proposed for the noise points generated after the clustering of the Tujia brocades. An experiment proved that the new method is remarkably effective for eliminating isolated noise points. Unsupervised clustering did not perform image segmentation semantically, so the clustered image needed post-processing to merge the clustered regions to form a whole segmentation object. Clustering-based image segmentation has high computational efficiency, but it is difficult to achieve image semantic segmentation because this method is based on low-level features of the image. In follow-up work, we plan to design an unsupervised image segmentation model by combining clustering with deep learning. It will use the feature extracted by a CNN for clustering, the clustering category labels as supervision information, and complete end-to-end Tujia brocade semantic segmentation.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S., et al. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Machine Intelligence* 34, 2274–2282. doi: 10.1109/TPAMI.2012.120

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach, 2nd Edn*. New York, NY: Springer-Verlag.

Chakrabarti, A., and Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *North-Holland* 7, 583–605. doi: 10.1016/B978-0-444-51862-0.50018-6

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Machine Intelligence* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Eur. Conf. Comp. Vis*. 2018:49. doi: 10.1007/978-3-030-01234-2_49

Fan, J., Yau, D. K. Y., Elmagarmid, A. K., and Aref, W. G. (2001). Automatic image segmentation by integrating color-edge extraction and seeded region growing. *IEEE Trans. Image Proces. Publ. IEEE Sign. Proces. Soc.* 10:1454. doi: 10.1109/83.951532

Felzenszwalb, P. F., and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *Int. J. Comp. Vis.* 59, 167–181. doi: 10.1023/B:VISI.0000022288.19776.77

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. *IEEE Trans. Pattern Anal. Machine Intelligence*. 2017:322. doi: 10.1109/ICCV.2017.322

Heath, M. D., Sarkar, S., Sanocki, T., and Bowyer, K. W. (1997). A robust visual method for assessing the relative performance of edge-detection algorithms. *IEEE Trans. Pattern Anal. Machine Intelligence* 19, 1338–1359. doi: 10.1109/34.643893

Jiang, G. M., Zhang, D., Cong, H. L., Zhang A. L., and Gao, Z. (2014). Automatic identification of jacquard warp-knitted fabric patterns based on the wavelet transform. *Fibr. Textil. Eastern Europe* 22, 53–56. doi: 10.3724/SP.J.1146.2010.00388

Kanezaki, A. (2018). "Unsupervised image segmentation by backpropagation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 1543–1547. doi: 10.1109/ICASSP.2018.8462533

Kuo, C.-F. J., and Shih, C. Y. (2011). Printed fabric computerized automatic color separating system. *Textile Res. J.* 81, 706–713. doi: 10.1177/0040517510383619

Kuo, C.-F. J., Shih, C. Y., and Lee, J. Y. (2005). Repeat pattern segmentation of printed fabrics by hough transform method. *Textile Res. J.* 75, 779–783. doi: 10.1177/0040517505058848

Kuo, C.-F. J., Su, T. L., and Huang, Y. J. (2007). Computerized color separation system for printed fabrics by using backward-propagation neural network. *Textile Res. J.* 8, 529–536. doi: 10.1007/BF02875876

Lachkar, A., Benslimane, R., D'Orazio, L., and Martuscelli, E. (2006). A system for textile design patterns retrieval. Part I: design patterns extraction by adaptive and efficient color image segmentation method. *J. Textile Instit. Fibre Sci. Textile Technol*. 97, 301–312. doi: 10.1533/joti.2005.0124

Liu, L., Wang, Q., Dong, M., Zhang, Z., Li, Y., Wang, Z., et al. (2020). "Application of K-means ++ algorithm based on t-SNE dimension reduction in transformer district clustering," in *2020 Asia Energy and Electrical Engineering Symposium* (Chengdu: AEEES), 74–78.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Machine Intelligence* 39, 640–651. doi: 10.1109/CVPR.2015.7298965

Luan, S., Kong, X., Wang, B., Guo, Y., and You, X. (2012). "Silhouette coefficient based approach on cell-phone classification for unknown source images," in *2012 IEEE International Conference on Communications* (Beijing: ICC), 6744–6747.

Marutho, D., Hendra Handaka, S., Wijaya, E., and Muljono. (2018). "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 International Seminar on Application for Technology of Information and Communication* (Semarang), 533–538. doi: 10.1109/ISEMANTIC.2018.8549751

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Machine Intelligence* 24, 971–987. doi: 10.1109/TPAMI.2002.1017623

Otsu, N. (2007). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybernet.* 9.1, 62–66. doi: 10.1109/TSMC.1979.4310076

Philipp, K., and Koltun, V. (2012). *Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials*. New York, NY: Curran Associates Inc.

Sulochana, S., and Vidhya, R. (2013). Texture based image retrieval using framelet transform–gray level co-occurrence matrix (GLCM). *Int. J. Adv. Res. Artificial Intelligence* 2:211. doi: 10.14569/IJARAI.2013.020211

Wan, Y., and Nie, L. (2018). The weaving and aesthetics of Tujia brocade patterns. *Fine Art Observ.* 11, 135–136. doi: 10.3969/j.issn.1006-8899.2018.11.036

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. (2017). Towards k-means-friendly spaces: simultaneous deep learning and clustering. *Proc. 34th Int. Conf. Machine Learn.* 70, 3861–3870.

Conflict of Interest: The 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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 26 October 2021
doi: 10.3389/fnbot.2021.752752

Fractional Wavelet-Based Generative Scattering Networks

Jiasong Wu [1,2,3,4]****, Xiang Qiu*** [1,4]***, Jing Zhang*** [1,4]***, Fuzhi Wu*** [1,4]***, Youyong Kong*** [1,2,4]***, Guanyu Yang*** [1,2,4]***, Lotfi Senhadji*** [3,4] ***and Huazhong Shu*** [1,2,4]

[1] Laboratory of Image Science and Technology, Key Laboratory of Computer Network and Information Integration, Southeast University, Ministry of Education, Nanjing, China, [2] Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing, China, [3] Univ Rennes, INSERM, LTSI-UMR 1099, Rennes, France, [4] Centre de Recherche en Information Biomédicale Sino-Français (CRIBs), Univ Rennes, INSERM, Rennes, France

Generative adversarial networks and variational autoencoders (VAEs) provide impressive image generation from Gaussian white noise, but both are difficult to train, since they need a generator (or encoder) and a discriminator (or decoder) to be trained simultaneously, which can easily lead to unstable training. To solve or alleviate these synchronous training problems of generative adversarial networks (GANs) and VAEs, researchers recently proposed generative scattering networks (GSNs), which use wavelet scattering networks (ScatNets) as the encoder to obtain features (or ScatNet embeddings) and convolutional neural networks (CNNs) as the decoder to generate an image. The advantage of GSNs is that the parameters of ScatNets do not need to be learned, while the disadvantage of GSNs is that their ability to obtain representations of ScatNets is slightly weaker than that of CNNs. In addition, the dimensionality reduction method of principal component analysis (PCA) can easily lead to overfitting in the training of GSNs and, therefore, affect the quality of generated images in the testing process. To further improve the quality of generated images while keeping the advantages of GSNs, this study proposes generative fractional scattering networks (GFRSNs), which use more expressive fractional wavelet scattering networks (FrScatNets), instead of ScatNets as the encoder to obtain features (or FrScatNet embeddings) and use similar CNNs of GSNs as the decoder to generate an image. Additionally, this study develops a new dimensionality reduction method named feature-map fusion (FMF) instead of performing PCA to better retain the information of FrScatNets,; it also discusses the effect of image fusion on the quality of the generated image. The experimental results obtained on the CIFAR-10 and CelebA datasets show that the proposed GFRSNs can lead to better generated images than the original GSNs on testing datasets. The experimental results of the proposed GFRSNs with deep convolutional GAN (DCGAN), progressive GAN (PGAN), and CycleGAN are also given.

Keywords: generative model, fractional wavelet scattering network, image generation, image fusion, feature-map fusion

OPEN ACCESS

Edited by:
Dongming Zhou,
Yunnan University, China

Reviewed by:
Jianhua Yin,
Shandong University, China
Gao Chen,
Dongguan University of Technology, China
Jinyu Cai,
Fuzhou University, China

****Correspondence:***
Jiasong Wu
jswu@seu.edu.cn

Received: *03 August 2021*
Accepted: *20 September 2021*
Published: *26 October 2021*

Citation:
Wu J, Qiu X, Zhang J, Wu F, Kong Y, Yang G, Senhadji L and Shu H (2021) Fractional Wavelet-Based Generative Scattering Networks. Front. Neurorobot. 15:752752. doi: 10.3389/fnbot.2021.752752

INTRODUCTION

Generative models have recently attracted the attention of many researchers, and they are widely used in image synthesis, image restoration, image inpainting, image reconstruction, and other applications. Many generative models have been proposed in the literature. They can be roughly classified into two types (Goodfellow et al., 2014): explicit density and implicit density models.

Among explicit density generative models, variational autoencoders (VAEs) (Kingma and Welling, 2014) and their variants (Rezende et al., 2014; Salimans et al., 2015; Gregor et al., 2018) are most likely the most commonly used models, since they have useful latent representation, which can be used in inference queries. Kingma and Welling (2014) were the first to propose VAEs, which train an encoder and decoder simultaneously and can perform efficient inference and learning in directed probabilistic models and in the presence of continuous latent variables with intractable posterior distributions. Salimans et al. (2015) bridged the gap between Markov chain Monte Carlo (MCMC) and VAEs, and incorporated one or more steps of MCMC into variational approximation. Sohn et al. (2015) proposed a conditional VAE (CVAE), which joins existing label information in training to generate corresponding category data. Rezende and Mohamed (2015) introduced a new approach for specifying flexible, arbitrarily complex, and scalable approximate posterior distributions and made a clear improvement in the performance and applicability of variational inference. Sønderby et al. (2016) presented a ladder variational autoencoder, which uses a process similar to a ladder network and recursively corrects the generation distribution based on a data-independent approximate likelihood. Higgins et al. (2017) presented a β-VAE, which is a modification of a variational autoencoder (VAE), with special emphasis on discovering disentangled latent factors. Oord et al. (2017) proposed a simple yet powerful generative model that learns discrete representations and allowed the model to circumvent issues of posterior collapse. Gregor et al. (2018) proposed temporal difference VAE (TD-VAE), which is a generative sequence model that learns representations containing explicit beliefs about states in several steps into the future Razavi et al. (2019) proposed vector quantized variational autoencoder (VQ-VAE), which augments with powerful priors over latent codes and is able to generate samples with a quality that rival those of state-of-the-art GANs on multifaceted datasets, such as ImageNet. Simonovsky and Komodakis (2018) proposed Graph VAE, sidesteps the hurdles of linearization of discrete structures by outputting a probabilistic fully connected graph of a predefined maximum size directly at once. For more references on VAEs, see Blei et al. (2017).

Among implicit density generative models, generative adversarial networks (GANs) (Goodfellow et al., 2014) and their variants (Chen et al., 2016; Radford et al., 2016) are probably the most commonly used models, since they provide better generated images than other generative models. Goodfellow et al. (2014) were the first to propose GANs, which estimate generative models *via* an adversarial process, where a generative model G and a discriminative model D are trained simultaneously without the need for Markov chains or unrolled approximate inference networks during either training or the generation of samples. However, the application of GANs to real-world computer vision problems still encounters at least three significant challenges (Wang et al., 2021): (1) high-quality image generation; (2) diverse image generation; and (3) stable training. Therefore, many variants of GANs have been proposed to handle the three challenges. The variants can be roughly classified into two groups (Wang et al., 2021): architecture variant GANs and loss variant GANs.

In terms of architecture variant GANs, for example, Radford et al. (2016) proposed deep convolutional GAN (DCGAN), which uses a convolutional neural network (CNN) as the discriminator D and deploys a deconvolutional neural network architecture for G; the spatial upsampling ability of the deconvolution operation enables the generation of images with higher resolution compared with the original GANs. Mirza and Osindero (2014) proposed conditional GAN (CGAN), which imposes a condition of additional information, such as a class label, to control the process of data generation in a supervised manner. Chen et al. (2016) presented InfoGAN, which decomposes an input noise vector into a standard incompressible latent vector and another latent variable to capture salient semantic features of real samples. Karras et al. (2018) presented progressive GAN (PGAN) for generative high-resolution images using the idea of progressive neural networks (Rusu et al., 2017), which does not suffer from forgetting and is able to deploy prior knowledge *via* lateral connections to previously learned features. Karras et al. (2020a,b) proposed StyleGAN, which leads to an automatically learned, unsupervised separation of high-level attributes and stochastic variation in generated images and, thus, enables intuitive, scale-specific control of the synthesis. More recently, Hudson and Zitnick (2021) introduced the Generative Adversarial Transformer (GANformer), which is a generalization of the StyleGAN and a simple yet effective generalization of the vanilla transformer, for a visual synthesis task.

In terms of loss-variant GANs, for example, Arjovsky et al. (2017) proposed Wasserstein GAN (WGAN), which uses the Wasserstein distance as the loss measure for optimization instead of Kullback–Leibler divergence. Gulrajani et al. (2017) proposed an improved method for training the discriminator for a WGAN, by penalizing the norm of discriminator gradients with respect to data samples during training rather than performing parameter clipping. Nowozin et al. (2016) proposed an alternative cost, which is a function of the f-divergence, for updating the generator, which is less likely to saturate at the beginning of training. Zhu et al. (2017) proposed CycleGAN for the task of image-to-image translation. Qi (2020) presented loss-sensitive GAN (LS-GAN), which trains the generator to produce realistic samples by minimizing the designated margins between real and generated samples. Miyato et al. (2018) proposed spectral normalization GAN (SN-GAN), which uses a weight normalization technique to train the discriminator more stably. Brock et al. (2019) proposed BigGAN, which uses hinge loss instead of Jensen–Shannon divergence and a large-scale dataset to train the generator to produce more realistic samples.

Although GANs and VAEs are great generative models, they raise many questions. A significant disadvantage of VAEs is that the resulting generative models produce blurred images compared with GANs, since the quality of VAEs crucially relies on the expressiveness of their inference models. A significant disadvantage of GANs is that the training process is very difficult and may lead to unstable training and model collapse. To design a network that can maintain the characteristics of high-quality generated images of GANs as much as possible while reducing the training difficulty of GANs, Angles and Mallat (2018) proposed generative scattering networks (GSNs), which use wavelet scattering networks (ScatNets) (Bruna and Mallat, 2013) as the encoder to obtain features (or ScatNet embeddings) and the deconvolutional neural network of DCGAN (Radford et al., 2016) as the decoder to generate an image. The advantage of GSNs is that there is no need to learn the parameters of ScatNets; therefore, the difficulty of training is reduced when compared with DCGAN, while the disadvantage of GSNs is that generated images can lose details, which affects the quality of the generated images. After careful inspection, we determined that the sources of relatively low-quality generated images of GSNs include at least two aspects: (1) the expression ability of ScatNets is slightly weaker than that of CNNs used in DCGAN; (2) applying PCA (Abdi and Williams, 2010) to reduce the dimension of the feature map of ScatNets in the encoder part of GSNs leads to an overfitting problem in the testing process of GSNs. This finding leads to the central question of our study:

Can we change the feature extraction method of ScatNets to a more powerful one that still does not need learning? Can we develop a more suitable dimensionality reduction method to solve the overfitting problem in the testing process of GSNs?

In an attempt to solve the above questions, in this study, we propose generative fractional scattering networks (GFRSNs), which can be seen as an extension of GSNs. The contributions of this article are as follows:

1) We use, for more expressiveness, fractional wavelet scattering networks (FrScatNets) (Liu et al., 2019) instead of ScatNets (Bruna and Mallat, 2013) to extract features of images, and we use image fusion (Liu et al., 2016; Yang et al., 2017) in GFRSNs to effectively improve the visual quality of the generated images.
2) We propose a new dimensionality reduction method named feature-map fusion (FMF), which is more suitable for reducing the feature dimension of FrScatNets than PCA, since the FMF method greatly alleviates the overfitting problem on the testing datasets using GFRSNs.
3) The image generated by the proposed GFRSN on the test set is better than that produced by the original GSNs.

The remainder of this article is organized as follows: In section Generative Scattering Networks (GSNs), wavelet scattering networks and the architectural components of GSNs are briefly introduced. The main architectural components of GFRSNs, which include fractional wavelet scattering networks, the FMF dimensionality reduction method and generative networks, and an image fusion method are introduced in section Generative Fractional Scattering Networks (GFRSNs). The performance of GFRSNs is analyzed and compared with that of the original GSNs in section Numerical Experiments. The conclusions and further discussion are presented in section Conclusions.

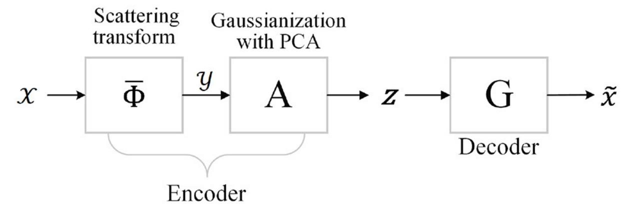


FIGURE 1 | Structure of generative scattering networks (GSNs) with principal component analysis (PCA) dimensionality reduction method.

GENERATIVE SCATTERING NETWORKS (GSNS)

In this section, we first briefly recall the generative scattering networks (GSNs) (Angles and Mallat, 2018), whose structure is shown in **Figure 1**.

The input Mth-order tensor $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times\cdots\times N_K}$, where $\mathbb{R}$denotes the real domain and each $N_i, i = 1, 2, 3, \cdots K$, addresses the i-mode of $\mathcal{X}$, and is first fed into the feature extraction part of the encoder to obtain the ScatNet features $y \in \mathbb{R}^{M_1\times M_2\times\cdots\times M_L}$. The next part of the encoder aims to map the features to a Gaussian latent variable $\mathbf{z} \in \mathbb{R}^U$, which is accomplished by whitening and projection to a lower-dimensional space. Inspired by Zou and Lerman (2019), we refer to this process as Gaussianization. Decoder G can be seen as a generator and is trained by minimizing the reconstruction loss between the output $\tilde{\mathcal{X}} \in \mathbb{R}^{N_1\times N_2\times\cdots\times N_K}$ and input $\mathcal{X}$. In other words, the generator calculation is regarded as the inverse problem of the scattering transform. The main components of GSNs include ScatNets, Gaussianization with PCA, and the generative network G. These components are recalled as follows.

Wavelet Scattering Networks (ScatNets)

Let the complex bandpass filter ψ_λ be constructed by scaling and rotating a filter ψ, respectively, by2^j and δ, as follows (Bruna and Mallat, 2013):

$$\psi_\lambda(t) = 2^{2j}\psi\left(2^j\delta^{-1}t\right), \lambda = 2^j\delta, \tag{1}$$

with $0 \le j \le J-1$, and $\delta = k\pi/K, k = 0, 1, ..., K-1$.

The wavelet-modulus coefficients of x are given by:

$$U[\lambda]x = |x * \psi_\lambda|. \tag{2}$$

The scattering propagator $U\left[p\right]$ is defined by cascading wavelet-modulus operators

$$\begin{aligned} U\left[p\right]x &= U[\lambda_m]\cdots U[\lambda_2]\,U[\lambda_1]\,x \\ &= \left|\left|\left|x * \psi_{\lambda_1}\right| * \psi_{\lambda_2}\right|\cdots * \psi_{\lambda_m}\right|, \end{aligned} \tag{3}$$

where $p = (\lambda_1, \lambda_2, .., \lambda_m)$ are the frequency-decreasing paths; in other words, $|\lambda_k| \geq |\lambda_{k+1}|, k = 1, 2, ..., m-1$. Note that $U[\varnothing]x = x$, and $\varnothing$ expresses the empty set.

The scattering operator S_J performs spatial averaging on a domain whose width is proportional to 2^J:

$$S[p]x = U[p]x * \phi_J = U[\lambda_m]\cdots U[\lambda_2]\,U[\lambda_1]\,x * \phi_J$$
$$= \left|\left||x * \psi_{\lambda_1}| * \psi_{\lambda_2}\right| \cdots * \psi_{\lambda_m}\right| * \phi_J. \quad (4)$$

The network nodes of layer m correspond to the set $P^m$of all paths $p = (\lambda_1, \lambda_2, .., \lambda_m)$ of length m. This m-th layer stores the propagated signals$\{U[p]x\}_{p\in P^m}$and outputs the scattering coefficients $\{S[p]x\}_{p\in P^m}$. The output is obtained by cascading the scattering coefficients of every layer.

Note that x in (2) can be one-dimensional data $\boldsymbol{x} \in \mathbb{R}^{N_1}$, two-dimensional data $\boldsymbol{X} \in \mathbb{R}^{N_1\times N_2}$, and third-order tensor $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times N_3}$, which can be seen as N_3 two-dimensional data $\boldsymbol{X} \in \mathbb{R}^{N_1\times N_2}$, and ScatNet addresses with these $\boldsymbol{X_s}$ one by one and then superimposes the results as output features. According to Mallat (2012), if we feed the input $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times N_3}$ into ScatNet, then we can obtain ScatNet features (or ScatNet embeddings) as follows:

$$y = S[p]\,\mathcal{X} \in \mathbb{R}^{N_3\times(1+LJ+L^2J(J-1)/2)\times(N_1/2^J)\times(N_2/2^J)}, \quad (5)$$

where N_3 is the number of input sample channels, and N_1 and N_2 are the width and height of the input sample, respectively. $N_1/2^J$ and $N_2/2^J$ are the width and height of the output features. J is a scale factor, and L is the number of rotation angles. Note that the number of feature maps in the first, second, and third layers is 1, LJ, and $L^2J(J$-1)/2, respectively.

Gaussianization With PCA

As shown in **Figure 1**, the last step of the encoder maps the transformed features in such a way that we can sample from the Gaussian distribution to generate new images, as required by the generator. Specifically, let $\{y\}_{t=1}^{T}$ be the output features of the ScatNet embedding, and let y be the representing matrix of $\{y\}_{t=1}^{T}$, while $\mathbf{z}$ is the latent variable of the generator. As advocated in Angles and Mallat (2018), $\mathbf{z}$ can be interpreted as an address, with a dimension d lower than that in the input image. Hence, to get a lower-dimensional embedding of the output features, one can perform principal component analysis (PCA) (Abdi and Williams, 2010) to project the features of the scattering transform to a lower-dimensional space.

Next, to whiten them, we choose $u = \frac{1}{T}\sum_{t=1}^{T} y, \sum = \frac{1}{T}\sum_{t=1}^{T}(y-u)(y-u)^*$, and the whitening map $A = \sum^{-1/2}(Id - u)$.

Hence, the resulting embedding of the encoder is

$$\mathbf{z} = \sum\nolimits^{-1/2}(y - u). \quad (6)$$

After the above process, the whitened sample is uncorrelated, and their distribution will be close to a normal one with identity covariance (Angles and Mallat, 2018), which is exactly what we want to feed to the generator.

Generator Networks in GSNs

The generative network G of GSNs is a neural one, which is similar to the generator of DCGAN (Radford et al., 2016), which inverts the whitened scattering embedding on training samples. The generator network G includes a fully connected layer (FC), batch normalization layer (BN) (Ioffe and Szegedy, 2015), bilinear upsampling (Upsample) layer, and convolutional layer (Conv2d) with a kernel size of 7 × 7. Except for the last layer, which uses the tanh activation function, the others use the default ReLU (Nair and Hinton, 2010) activation function.

Generative scattering networks with PCA as the dimensionality reductional method choose the L_1-norm loss function and solve the following optimization problem (Zhao et al., 2017):

$$g_1 = min\,Loss_{L_1}(\mathcal{X}, \tilde{\mathcal{X}}) = min\,\frac{1}{N}\sum_{i=1}^{N}\left|\mathcal{X}^{(i)} - \tilde{\mathcal{X}}^{(i)}\right|, \quad (7)$$

where $\mathcal{X}$ represents the input data, $\tilde{\mathcal{X}}$ represents the generative data, $\mathcal{X}^{(i)}$ represents the i-th input sample, and $\tilde{\mathcal{X}}^{(i)}$ represents the i-th generative sample:

$$\tilde{\mathcal{X}} = G(PCA(S[p]\,\mathcal{X})), \quad (8)$$

where $S[p]\mathcal{X}$ denotes the feature extraction process with ScatNets, and PCA(.) represents that the feature dimensionality reduction method is PCA. G(.) represents the generative network G. The optimization problems in (7) are then solved with the Adam optimizer (Kingma and Ba, 2015) using the default hyperparameters.

GENERATIVE FRACTIONAL SCATTERING NETWORKS (GFRSNS)

In this section, we introduce the proposed generative fractional scattering networks (GFRSNs), whose structure is shown in **Figure 2**.

The input $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times\cdots\times N_K}$ is first fed into the fractional wavelet scattering networks (FrScatNets) to obtain FrScatNet features (or FrScatNet embeddings) $y_\alpha \in \mathbb{R}^{M_1\times M_2\times\cdots\times M_L}$, whose dimensions are then reduced by the proposed feature-map fusion (FMF) method to obtain an implicit tensor $z_\alpha \in$

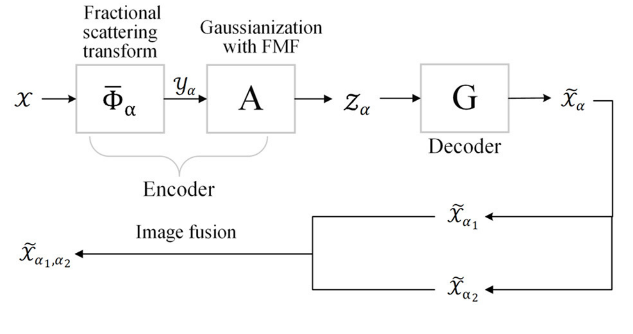


FIGURE 2 | Structure of generative fractional scattering networks (GFRSNs).

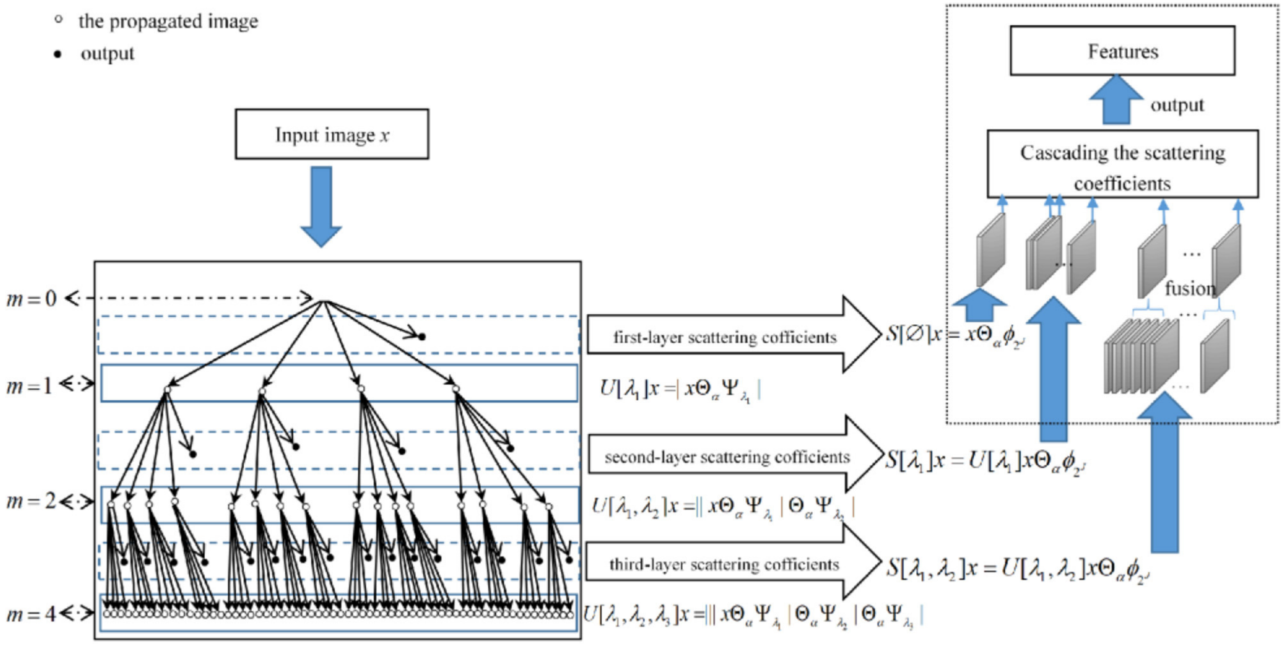

the propagated image
output
Input image x
$m=0$
$m=1$
$m=2$
$m=4$
first-layer scattering cofficients
$U[\lambda_1]x = |x\Theta_\alpha \Psi_{\lambda_1}|$
second-layer scattering cofficients
$U[\lambda_1,\lambda_2]x = ||x\Theta_\alpha \Psi_{\lambda_1}|\Theta_\alpha \Psi_{\lambda_2}|$
third-layer scattering cofficients
$U[\lambda_1,\lambda_2,\lambda_3]x = |||x\Theta_\alpha \Psi_{\lambda_1}|\Theta_\alpha \Psi_{\lambda_2}|\Theta_\alpha \Psi_{\lambda_3}|$
$S[\varnothing]x = x\Theta_\alpha \phi_{2^J}$
$S[\lambda_1]x = U[\lambda_1]x\Theta_\alpha \phi_{2^J}$
$S[\lambda_1,\lambda_2]x = U[\lambda_1,\lambda_2]x\Theta_\alpha \phi_{2^J}$
Features
output
Cascading the scattering coefficients
fusion
...

FIGURE 3 | Fractional wavelet scattering network and the feature-map fusion dimensional reduction method.

$\mathbb{R}^{O_1\times O_2\times\cdots\times O_K}$, which is then fed into the generator G to obtain the generated output tensor $\tilde{\mathcal{X}}_\alpha \in \mathbb{R}^{N_1\times N_2\times\cdots\times N_K}$. In other words, the generative network G is seen as the inverse problem of FrScatNets. The main components of GFRSNs include FrScatNets, Gaussianization with feature-map fusion dimensionality reduction method, and the generative network G. In the following, these components of GFRSNs are introduced.

Fractional Wavelet Scattering Networks (FrScatNets)

In this subsection, fractional wavelet scattering networks (FrScatNets) (Liu et al., 2019) are briefly introduced. Similar to (2), the fractional wavelet modulus coefficients of x are given by:

$$U_\alpha [\lambda]\, x = |x\Theta_\alpha \psi_\lambda|\,, \tag{9}$$

where Θ_α is the fractional convolution defined by Shi et al. (2010);

$$x(t)\,\Theta_\alpha \psi_\lambda (t) = e^{-\frac{j}{2}t^2\cot\theta}\left[\left(x(t)\,e^{\frac{j}{2}t^2\cot\theta}\right)*\psi_\lambda (t)\right], \tag{10}$$

where the parameter α is the fractional order and $\theta = \alpha\pi/2$ represents the rotation angle. Note that when $\alpha = 1$, the fractional convolution in (10) reduces to conventional convolution in (2).

The fractional scattering propagator $U_\alpha\left[p\right]$ is defined by cascading fractional wavelet modulus operators

$$\begin{aligned} U_\alpha\left[p\right]x &= U_\alpha\left[\lambda_m\right]\cdots U_\alpha\left[\lambda_2\right]U_\alpha\left[\lambda_1\right]x \\ &= \left|\left|\,\left|x\Theta_\alpha\psi_{\lambda_1}\right|\Theta_\alpha\psi_{\lambda_2}\right|\cdots\Theta_\alpha\psi_{\lambda_m}\right|, \end{aligned} \tag{11}$$

where $p = (\lambda_1, \lambda_2, .., \lambda_m)$are the frequency-decreasing paths; in other words, $|\lambda_k| \geq \left|\lambda_{k+1}\right|$, $k = 1, 2, ..., m-1$. Note that $U_\alpha\left[\varnothing\right]x = x$, and $\varnothing$ expresses the empty set.

The fractional scattering operator S_α performs spatial averaging on a domain whose width is proportional to 2^J:

$$\begin{aligned} S_\alpha\left[p\right]x &= U_\alpha\left[p\right]x\Theta_\alpha\phi_J = U_\alpha\left[\lambda_m\right]\cdots U_\alpha\left[\lambda_1\right]x\Theta_\alpha\phi_J \\ &= \left|\left|\,\left|x\Theta_\alpha\psi_{\lambda_1}\right|\Theta_\alpha\psi_{\lambda_2}\right|\cdots\Theta_\alpha\psi_{\lambda_m}\right|\Theta_\alpha\phi_J. \end{aligned} \tag{12}$$

The structure of FrScatNets is shown on the left of **Figure 3**.

The network nodes of the layer m correspond to the set P^m of all paths $p = (\lambda_1, \lambda_2, .., \lambda_m)$ of length m. This m-th layer stores the propagated signals $\left\{U_\alpha\left[p\right]x\right\}_{p\in P^m}$ and outputs the fractional scattering coefficients $\left\{S_\alpha\left[p\right]x\right\}_{p\in P^m}$. The output is obtained by cascading the fractional scattering coefficients of every layer. Note that when $\alpha = 1$, the FrScatNets in (12) default to conventional ScatNets in (4), since the fractional convolution in (10) reduces to conventional convolution in (2).

Note that FrScatNets retain the advantages of ScatNets, for example, no need for learning, translation-invariant property, linearized deformations, and certain parameters. Compared with ScatNets, FrScatNet adds a free parameter α, which represents fractional order. With α continuously growing from 0 to 2, FrScatNets can show the characteristics of an image from time domain to frequency domain. Thus, FrScatNets provide more fractional domain choices for the feature extraction of input data. Furthermore, for the image generation task in this study, we can obtain as many generated images from FrScatNet embeddings as different fractional orders α_i, and then they can be fused to further improve the quality of the generated images.

If we feed the input $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times N_3}$ into the FrScatNet, then we can obtain the features of FrScatNet (or FrScatNet embeddings) as follows:

$$y_\alpha = S_\alpha\left[p\right] X \in \mathbb{R}^{N_3\times\left(1+LJ+L^2J(J-1)/2\right)\times\left(N_1/2^J\right)\times\left(N_2/2^J\right)}. \quad (13)$$

Note that the size of output features of FrScatNets is the same as that of ScatNets, whose size is shown in (5).

Gaussianization With FMF

In this subsection, we introduce a new method called FMF to reduce the dimensionality of the features after a fractional scattering transformation. We propose such an algorithm based on the hierarchical tree structure of features extracted by the fractional scattering transform to replace PCA to map the features to a low-dimensional space. More specifically, since the output features of different layers from the fractional scattering transform have a hierarchical structure, which is not considered in the PCA algorithm, we need a dimensionality reduction method that can make full use of this hierarchical information. The number of feature maps in the first, second, and third layers is 1, *LJ*, and L^2*J(J-1)/2*, respectively. Obviously, the third layer has the largest number of feature maps. Therefore, we fuse only the feature maps from the third layer of the fractional scattering transform to significantly reduce the data dimension. The fusion method is very simple: we obtain a new feature map by simply taking the average of every *L(J-1)/2* feature map, which obtains *LJ* feature maps after applying the FMF method to the output of the third layer of FrScatNets. The dotted box on **Figure 3** illustrates the proposed FMF method.

Therefore, an input tensor $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times N_3}$is fed into the FrScatNets to obtain FrScatNet features y_α in (13), which are then processed by the FMF method, obtaining an implicit tensor

$$z_\alpha = FMF\left(y_\alpha\right) \in \mathbb{R}^{N_3\times(1+LJ+LJ)\times\left(N_1/2^J\right)\times\left(N_2/2^J\right)}, \quad (14)$$

whose size is significantly smaller than the size shown in (13) without using the FMF method. Note that FMF(.) means performing the FMF method.

The obtained implicit tensor z_α is then input to the generator network G, described below, to obtain the generated image.

Generative Networks in GFRSNs

The generative network G of GFRSNs is also a deconvolutional neural network that has a generator similar to that of DCGAN (Radford et al., 2016), which inverts fractional scattering embeddings on training samples. The generative network G of GFRSNs also includes a fully convolutional layer (Fully Conv) (Long et al., 2015) and several convolution blocks that consist of bilinear upsampling (UP), two convolutional layers (Conv) with a 3 × 3 kernel size, batch normalization, and ReLU (the activation function of the last convolution layer is tanh). GFRSNs also choose the L_1-norm loss function and solve the following optimization problem:

$$g_2 = \min\ Loss_{L_1}\left(\mathcal{X}, \tilde{\mathcal{X}}_\alpha\right) = \min \frac{1}{N}\sum_{i=1}^{N}\left|\mathcal{X}^{(i)} - \tilde{\mathcal{X}}_\alpha^{(i)}\right|, \quad (15)$$

where $\tilde{\mathcal{X}}_\alpha$ represents the generative data and $\tilde{\mathcal{X}}_\alpha^{(i)}$. represents the *i*-th generative sample, and

$$\tilde{\mathcal{X}}_\alpha = G\left(FMF\left(S_\alpha\left[p\right]\mathcal{X}\right)\right), \quad (16)$$

where $S_\alpha[p]\mathcal{X}$ denotes the feature extraction process with FrScatNets, FMF(.) represents the dimensionality reduction process, and $G(.)$ represents the generative network.

The optimization problem in (15) is then solved with the Adam optimizer (Kingma and Ba, 2015).

Image Fusion

In contrast to GSNs, the proposed generative fractional scattering networks (GFRSNs) embed the input using FrScatNets, which allows for deriving many embeddings, since FrScatNets have an additional fractional order α; therefore, we can embed the input in different fractional order domains. These FrScatNet embeddings may extract many different but complementary features from the input. We can effectively use these embeddings to generate many images and further improve the quality of the synthesized images using fusion methods. In this study, as shown at the bottom of **Figure 2**, we use a simple image fusion method that is weighted average. As examples, we simply use the following:

$$\tilde{\mathcal{X}}_{\alpha_1,\alpha_2} = \lambda\tilde{\mathcal{X}}_{\alpha_1} + (1-\lambda)\,\tilde{\mathcal{X}}_{\alpha_2}, \quad (17)$$

where λ is the balanced parameter, which is set here to 0.5.

NUMERICAL EXPERIMENTS

In this section, we evaluate the quality of the generated images by the proposed GFRSNs by means of several experiments. The quality of the generated images is evaluated with two criteria: peak signal to noise ratio (PSNR) (Wang et al., 2003) and structural similarity (SSIM) (Wang et al., 2004).

We performed experiments on two datasets that have different levels of variability: CIFAR-10 (Krizhevsky, 2009) and CelebA (Liu et al., 2015). The CIFAR-10 dataset includes 50,000 training images and 10,000 testing images, whose sizes are 32 × 32 × 3. In all the experiments on the CIFAR-10 dataset, after image grayscale preprocessing, the number of rotation angles L is set to 8, and the fractional scattering averaging scale is set to $2^J = 2^3 =$ 8, which means that we linearize translations and deformations of up to 8 pixels. Therefore, the size of the output features from FrScatNets according to Equation (13) is 1 × 217 × 4 × 4, which is then, after the FMF method according to Equation (14), reduced to 1 × 49 × 4 × 4 (the size of implicit tensor z_α). In addition, the CelebA dataset contains thousands of images, and we choose 65,536 training images and 16,384 testing images, whose sizes are 128 × 12 8 × 3. In all the experiments on the CelebA dataset, after image grayscale preprocessing, the number of rotation angle L is set to 8, and the fractional scattering averaging scale is set to $2^J = 2^4 = 16$, which means that we linearize translations and deformations of up to 16 pixels. Thus, the size of the output features from FrScatNets according to (13) is 1 × 417 × 8 × 8, which is then, after FMF method according

TABLE 1 | Core parameters of FrScatNet with and without feature dimensionality reduction.

Parameter	Descriptions	Dataset	
		CIFAR-10	**CelebA**
$N_1 \times N_2 \times N_3$	The size of input image	32 × 32 × 1	128 × 128 × 1
J	The fractional scattering averaging scale	3	4
L	The number of rotation angle	8	8
$\mathrm{N}_3\times(1+\mathrm{LJ}+\frac{\mathrm{L}^2\mathrm{J}(\mathrm{J}-1)}{2})\times\frac{\mathrm{N}_1}{2^\mathrm{J}}\times\frac{\mathrm{N}_2}{2^\mathrm{J}}$	The shape of FrScatNets features $\mathcal{Y}_\alpha$	1 × 217 × 4 × 4	1 × 417 × 8 × 8
$N_3 \times (1+2\times LJ)\times \frac{N_1}{2^J}\times\frac{N_2}{2^J}$	The shape of implicit tensor $\mathcal{Z}_\alpha$ with FMF	1×49×4×4	1×65×8×8

TABLE 2 | Peak signal to noise ratio (PSNR) and structural similarity (SSIM) scores of training and testing images from FrScatNets with fractional orders $\alpha_1 = \alpha_2 = 1$ on the CIFAR-10 dataset.

	PCA	**Feature-Map Fusion**	**Increased (%)**
Train PSNR	**23.08**	20.1500	−12.69
Test PSNR	17.92	**18.1000**	**1.00**
Train SSIM	**0.9428**	0.8859	−6.08
Test SSIM	0.8206	**0.8352**	**1.78**

Increased means the percentages of relative improvements of FMF over principal component analysis (PCA), the better results are shown in bold.

TABLE 3 | PSNR and SSIM scores of training and testing images from FrScatNets with fractional orders $\alpha_1 = \alpha_2 = 1$ on the CelebA dataset.

	PCA	**Feature-Map Fusion**	**Increased (%)**
Train PSNR	**23.8124**	22.7526	−4.45
Test PSNR	**20.5312**	19.7688	−3.71
Train SSIM	**0.9529**	0.944	−0.93
Test SSIM	**0.9104**	0.8993	−1.22

Increased means the percentage of relative improvements of FMF over PCA, the better results are shown in bold.

to Equation (14), reduced to 1 × 65 × 8 × 8 (the size of implicit tensor $\mathcal{Z}_\alpha$). **Table 1** shows the core parameters of FrScatNet and its settings on the CIFAR-10 and CelebA datasets.

In the following, we first compare the visual quality of the generated images with different feature dimensionality reduction methods in the framework of GFRSNs. Then, we compare the visual quality of the generated images with FrScatNets. Finally, we compare the visual quality of the fused images and unfused images. The following experiments are implemented using PyTorch on a PC machine, which sets up an Ubuntu 16.04 operating system and has an Intel (R) Core(TM) i7-8700K CPU with a speed of 3.7 GHz and 32 GB RAM, and has two NVIDIA GeForce GTX1080-Ti GPUs.

Image Generative Results With Different Dimensionality Reduction Methods

In this subsection, we compare the results on the quality of generative images with two different dimensionality reduction methods: the PCA method and the proposed FMF method. We set the fractional orders to be $\alpha_1 = \alpha_2 = 1$, and use conventional ScatNets to extract features from the input $\mathcal{X}$ for simplicity.

For the PCA-based GFRSNs, the flow chart is shown in **Figure 1**. For the CIFAR-10 dataset, the size of the implicit vector **z** is 49 × 4 × 4 = 784, and for the CelebA dataset, the size of the implicit vector **z** is 65 × 8 × 8 = 4,160. We use the PyTorch code of generative scattering networks[1] provided by Tomás Angles. The PSNR and SSIM on the CIFAR-10 and CelebA datasets are shown in the second columns of **Tables 2**, **3**, respectively.

As shown in the two tables, the scores of PSNR (Train PSNR) and SSIM (Train SSIM), both in the training dataset, are very good for the PCA-based GFRSNs; however, their corresponding values (test PSNR and test SSIM) in the testing dataset are slightly low. This phenomenon indicates that an overfitting problem has occurred using the PCA-based GFRSNs. We argue the reason behind this phenomenon could be that the output feature of FrScatNets $\mathcal{Y}_\alpha$ in (16) is a 4th-order tensor, which is performed by PCA to obtain an implicit vector **z**. This process loses correlations between various dimensions of the data. Therefore, we consider using FMF as the dimensionality reduction method to maintain the structures of the input data better.

For the proposed FMF-based GFRSNs, the flow chart is shown in **Figure 2**. The size of the implicit tensor $\mathcal{Z}_{\alpha_i}$ is 1 × 49 × 4 × 4 on CIFAR-10, and for the CelebA dataset, the size of implicit tensor $\mathcal{Z}_{\alpha_i}$ is 1 × 65 × 8 × 8. The PSNR and SSIM on the CIFAR-10 and CelebA datasets are shown in the third columns of **Tables 2**, **3**, respectively. As can be seen from the two tables, train PSNR and train SSIM of the FMF-based GFRSNs are slightly worse than those of the PCA-based GFRSNs on the CIFAR-10 and CelebA datasets; however, the test PSNR and test SSIM of the proposed FMF-based GFRSNs are better than those of the PCA-based GFRSNs. For example, Test PSNR and Test SSIM have relatively increased by 1 and 1.8%, respectively, when compared with the PCA-based GFRSNs, on the CIFAR-10 dataset. However, with regard to the CelebA dataset, Test PSNR and Test SSIM have decreased by 3.71 and 1.22%, respectively, when compared with the PCA-based GFRSNs. Nevertheless, the experimental results still show that the overfitting problem on the testing datasets can be alleviated with the FMF dimensionality reduction method.

Although the performance of the proposed FMF method on theCIFAR-10 dataset is better than that of PCA and has a similar generation ability on the CelebA dataset, more importantly, FMF has better generalization performance under the framework of GFRSNs. In other words, our generative model will not overfit on the test set. However, in order to better reflect the role of fractional scattering transformation and, hence, abolish the influence of FMF, we still use the PCA method in the following two experiments.

[1] https://github.com/tomas-angles/generative-scattering-networks

TABLE 4 | Results with FrScatNets on the CIFAR-10 dataset.

(α_1, α_2)	**Fusion**	**Test PSNR**	**Increased (%)**	**Test SSIM**	**Increased (%)**
Base line with $(\alpha_1, \alpha_2) =$ (1.00,1.00)	No	18.1000	0	0.8352	0
(0.10,1.00)	No	13.9738	−22.80	0.5442	−34.84
	Yes	16.9597	−6.30	0.6974	−16.50
(0.40,1.00)	No	**18.8280**	4.02	**0.8514**	**1.94**
	Yes	**18.9869**	4.90	**0.8970**	**7.40**
(0.70,1.00)	No	18.6614	3.10	0.8469	1.40
	Yes	18.8421	4.10	0.8887	6.40
(1.30,1.00)	No	18.6169	2.86	0.8462	1.32
	Yes	18.8059	3.90	0.8870	6.20
(1.60,1.00)	No	**18.8209**	**3.98**	**0.8517**	**1.98**
	Yes	**18.9688**	**4.80**	**0.8987**	**7.60**
(1.90,1.00)	No	14.0110	−22.59	0.5474	−34.46
	Yes	16.9959	−6.10	0.7041	−15.70
(1.00,0.10)	No	14.0099	−22.60	0.5498	−34.17
	Yes	16.9054	−6.60	0.6941	−16.90
(1.00,0.40)	No	18.9351	4.61	0.8550	2.37
	Yes	18.9507	4.70	0.8978	7.50
(1.00,0.70)	No	18.7289	3.47	0.8499	1.76
	Yes	18.7335	3.50	0.8753	4.80
(1.00,1.30)	No	18.6947	3.29	0.8434	0.98
	Yes	18.5887	2.70	0.8753	4.80
(1.00,1.60)	No	18.9056	4.45	0.8545	2.31
	Yes	18.9507	4.70	0.8987	7.60
(1.00,1.90)	No	14.0487	−22.38	0.5520	−33.91
	Yes	16.9778	−6.20	0.7074	−15.30
Fusing (0.40,1.00) and (1.60,1.00)	Yes	**19.1589**	**5.85**	**0.8927**	**6.89**

Some better results are shown in bold. In the second column, "No" means un-fused image, and "Yes" means fused image. We also show the percentages of relative improvements on Test PSNR and Test SSIM of FrScatNets of various fractional orders ($\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$) over the conventional ScatNets (the first row), respectively.

TABLE 5 | Results with FrScatNets on CelebA dataset.

(α_1, α_2)	**Fusion**	**Test PSNR**	**Increased (%)**	**Test SSIM**	**Increased (%)**
Base line with $(\alpha_1, \alpha_2) =$ (1.00,1.00)	No	21.1668	0	0.9221	0
(0.10,1.00)	No	18.3728	−13.2	0.7709	−16.4
	Yes	21.0186	−0.7	0.9156	−0.7
(0.40,1.00)	No	21.4631	**1.1**	**0.9350**	**3.3**
	Yes	22.2040	**5.3**	**0.9608**	**6.5**
(0.70,1.00)	No	21.3996	1.1	0.9525	2.3
	Yes	22.3098	5.4	0.9820	6.2
(1.30,1.00)	No	21.3785	1	0.9433	2.4
	Yes	22.3098	5.4	0.9793	6.2
(1.60,1.00)	No	**21.4631**	**1.4**	**0.9571**	**3.8**
	Yes	**22.3310**	**5.5**	**0.9839**	**6.7**
(1.90,1.00)	No	18.6268	−12	0.7866	−14.7
	Yes	21.1456	−0.1	0.9219	−0.02
(1.00,0.10)	No	18.2458	−13.8	0.7561	−18
	Yes	20.9551	−1	0.9092	−1.4
(1.00,0.40)	No	21.5055	1.6	0.9405	2
	Yes	22.3098	5.4	0.9756	5.8
(1.00,0.70)	No	21.2515	0.4	0.9249	0.3
	Yes	22.2251	5	0.9700	5.2
(1.00,1.30)	No	21.2303	0.3	0.9267	0.5
	Yes	22.2251	5	0.9581	3.9
(1.00,1.60)	No	21.4843	1.5	0.9433	2.3
	Yes	22.3098	5.4	0.9765	5.9
(1.00,1.90)	No	18.7115	−11.6	0.7912	−14.2
	Yes	21.1732	0.03	0.9212	−0.1
Fusing (0.40,1.00) and (1.60,1.00)	Yes	**22.0770**	**4.3**	**0.9802**	**6.3**

Some better results are shown in bold. In the second column, "No" means un-fused image, and "Yes" means fused image. We also show the percentages of relative improvements on Test PSNR and Test SSIM of FrScatNets of various fractional orders (α_1, α_2) over the conventional ScatNets (the first row), respectively.

Image Generative Results With Different Fractional Order α

In this subsection, we explore the impact of fractional order α on the quality of the generated image using the framework of GFRSNs shown in **Figure 1**. The other parameter settings of FrScatNets are shown in **Table 1**. We choose the L_1 loss function in (15) and train the generator with the Adam optimizer using the default hyperparameters.

In this subsection, we use a two-dimensional fractional Morlet wavelet to construct the FrScatNets. For the two-dimensional fractional wavelet, two fractional orders, α_1 and α_2, are needed to determine the rotational angle. The angle is defined as $\theta = \alpha\pi/2$, ranging from 0 to π; thus, the fractional orders α_1 and α_2 change from 0 to 2. To save computation time, we fix one order as 1 and the other order changes within the range 0–2 for computing the fractional scattering coefficients. The chosen values are 0.1, 0.4, 0.7, 1, 1.3, 1.6, and 1.9. The above parameter settings are same as those in Liu et al. (2016). Note that FrScatNets reduce to conventional ScatNets when $\alpha_1 = \alpha_2 = 1$. The PSNR and SSIM of the generated images from FrScatNets on the CIFAR-10 and CelebA datasets are shown in **Tables 4**, **5**.

Generally, as shown in **Table 4**, best results are not obtained using FrScatNets with $(\alpha_1, \alpha_2) = (1, 1)$, which means that FrScatNets with some fractional order choice of (α_1, α_2) obtain better embeddings than the conventional ScatNets. For example, both the PSNR and SSIM results are very good the FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1.00)$ were used and whose Test PSNR and Test SSIM increased by 4.2 and 1.9%, respectively, compared with those of the ScatNets.

For the CelebA dataset, as shown in **Table 5**, both the PSNR and SSIM scores in the test set are also very good when FrScatNets with $(\alpha_1, \alpha_2) = (1.6, 1)$ are used. Indeed, Test PSNR and Test SSIM increased by 1.4 and 3.8%, respectively, compared with those of the ScatNets.

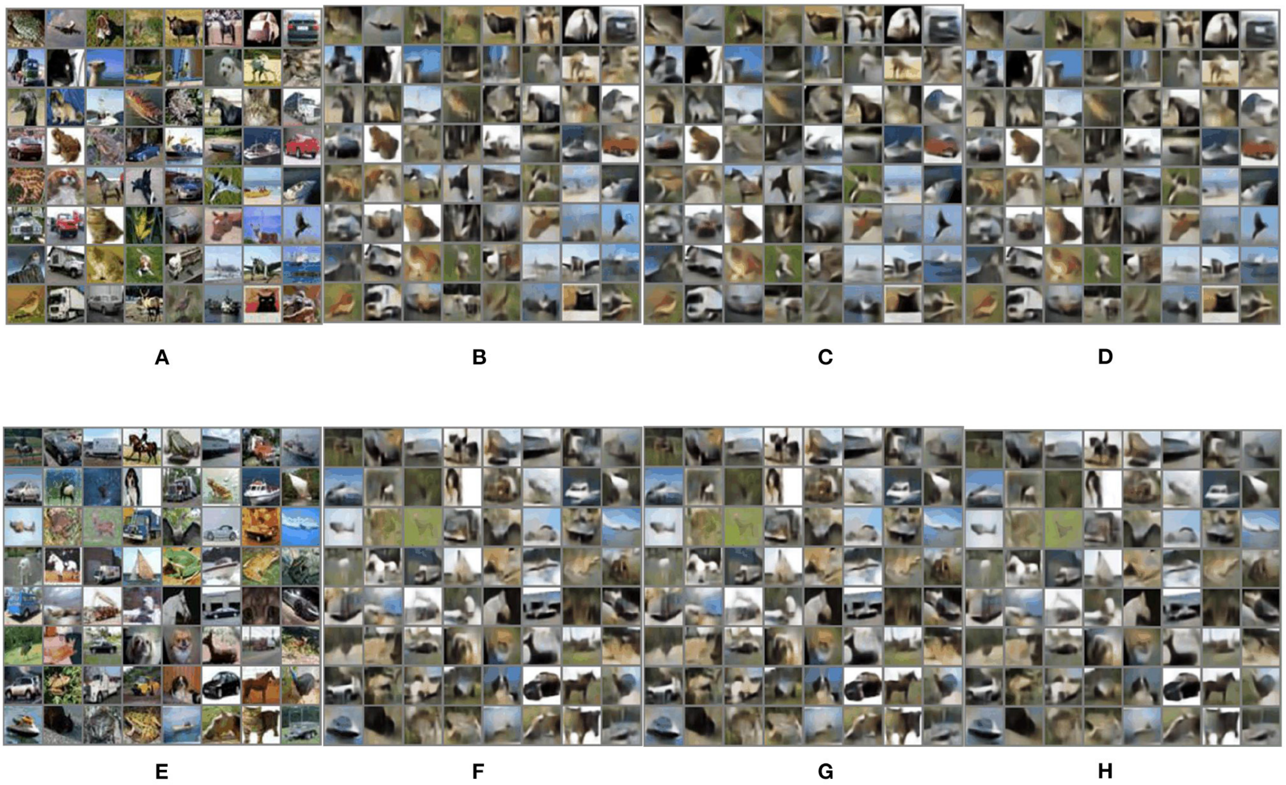


FIGURE 4 | Generative images on CIFAR-10 dataset using FrScatNet embeddings. **(A)** Original training images; **(B)** generative training images using FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1)$; **(C)** generative training images using FrScatNets with $(\alpha_1, \alpha_2) = (1, 1)$; **(D)** fused training image using FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1$ and $(\alpha_1, \alpha_2) = (1, 1)$; **(E)** original testing images; **(F)** generative testing images using FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1)$; **(G)** generative testing images using FrScatNets with $(\alpha_1, \alpha_2) = (1, 1)$; **(H)** fused testing image using FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1)$ and $(\alpha_1, \alpha_2) = (1, 1)$.

The generative images on the CIFAR-10 dataset using FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1)$ and $(\alpha_1, \alpha_2) = (1, 1)$ are shown in **Figure 4**. The generative images on the CelebA dataset using FrScatNets with $(\alpha_1, \alpha_2) = (1.6, 1)$ and $(\alpha_1, \alpha_2) = (1, 1)$ are shown in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

Image Generative Results With Image Fusion

In this subsection, we explore the impact of image fusion on the quality of the generated images using the framework of GFRSNs shown in **Figure 2**.

Since conventional GSNs are a good baseline for the framework of GFRSNs with different fractional orders (α_1, α_2), as an example, we consider the case in which the generative images from FrScatNets with different fractional orders (α_1, α_2), where α_1 and α_2 are not simultaneously equal to 1.00, are fused with the generative images from conventional ScatNets, in other words, FrScatNets with fractional orders $(\alpha_1, \alpha_2) = (1, 1)$. Since the fractional parameters can have multiple choices, naturally, we hope to explore the effect of image fusion under different fractional parameters. All the fused images are achieved using the average method shown in Equation (17), and we choose $\lambda = 0.5$. The PSNR and SSIM results of fused images on the CIFAR-10 dataset are shown in **Table 4**, and those on the CelebA dataset are shown in **Table 5**. Note that the results are shown in the row where the "Fusion or not?" column is "Yes" in **Tables 4**, **5**. As can be seen from the two tables, the results of PSNR and SSIM for the fused images are generally better than those for the unfused images from FrScatNets with different fractional orders (α_1, α_2), where α_1 and α_2 are not 1 at the same time. For example, when the generative images from FrScatNets with $(\alpha_1, \alpha_2) = (0.4, 1)$ are fused with the generative images from ScatNets, the Test PSNR and Test SSIM are increased from 18.828 and 0.8514 to 18.9869 and.897, respectively, on the CIFAR-10 dataset. The results are also better than those of ScatNet-based GFRSRNs, whose Test PSNR and Test SSIM are 18.1 and 0.8352, respectively. When the generative images from FrScatNets with $(\alpha_1, \alpha_2) = (1.6, 1)$ are fused with the generative images from ScatNets, the test PSNR and test SSIM are increased from 21.4632 and 0.9571 to 22.337 and 0.9839, respectively, on the CelebA dataset. The results are also better than those of ScatNet-based GFRSRNs, whose test PSNR and test SSIM are 21.1668 and 0.944, respectively. The fused images on the CIFAR-10 dataset are shown in **Figures 4D,H** and those on the CelebA dataset are shown in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, respectively.

We also consider the generative images from FrScatNets with fractional orders (0.4, 1) and (1.6, 1), and the results are shown in the last row of **Tables 4**, **5**, respectively. As can be seen from the two tables, the test PSNR and test SSIM are better than the fusion results of fractional orders (1.6, 1) and (1, 1) on both the CIFAR-10 and CelebA datasets.

The Deformation Property of the Proposed GFRSNs

In this section, we evaluate the deformation property of the proposed GFRSNs as generally done in GANs. Specifically, given two images x_1 and x_2, we modify β to get the interpolated images:

$$x_\beta = G\left((1-\beta)\, z_1 + \beta z_2\right), \text{for} z_1 = \Phi\left(x_1\right) \text{ and } z_2 = \Phi\left(x_2\right), \tag{18}$$

where Φ(.) denotes the fixed embedding, that is, the fractional scattering transform and Gaussianization process. The results are shown in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

As Angles and Mallat (2018) point out, the Lipschitz continuity to deformations of the scattering network resulting in the continuous deformation from one image to another image. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html shows that the proposed GFRSNs improve the capability to extract information while maintaining the deformation properties when compared with GSNs. On the other hand, we reproduce the morphing properties of GANs without learning a discriminator.

Besides, we should note that the generated images have strong similarities with those in the training set and, thus, lead to some unrealistic results; this is partially due to the autoencoder architecture of our model. Although under the autoencoder architecture, regarding the generative model as an inverse problem of FrScatNets, can eliminate our need to train an encoder or a discriminator, however, within this supervised paradigm, the generalization ability of the model may be limited to some extent. Therefore, when we try to recover images from unknown images, the results of the model will generate images that are similar to ones in the training set.

Comparison Results With GANs

In this section, we compared the results of the proposed GFRSNs with GANs on the CelebA dataset.

Comparison Results With DCGAN and PGAN

We compare the visual results of the proposed GFRSNs with those of the DCGAN (Radford et al., 2016) and progressive GAN (PGAN)[2] (Karras et al., 2018), as shown in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, from which we can see that DCGAN produces a certain degree of distortion. On the contrary, the proposed GFRSNs and PGAN do not show this kind of problem. PGAN generates more image details than the proposed GFRSNs, and we think that the reasons are:

(1) The proposed GFRSNs still belong to the autoencoder architecture, which is generally inferior to that of the GANs in terms of image generation quality. However, the autoencoder has its own merits; for example, it can obtain an image code (or a latent vector), which is very helpful for downstream tasks such as image classification. In contrast, the GANs cannot generate this latent vector.

(2) The proposed GFRSNs use learning-free FrScatNets instead of CNNs in the encoder stage, which significantly reduces the parameters (for example, reducing the parameters by half compared with DCGAN). However, it also has a certain impact on image generation performance.

(3) The proposed GFRSNs can maintain the structure of the face but show smoothed results to a certain extent. The reason for this is, maybe, the choice of L_1 loss.

(4) PGAN uses a more advanced low-resolution to high-resolution generation paradigm, which is more effective than the generator used in GFRSNs.

Note that we choose DCGAN as one of the compared methods, since we use the same generator architecture as the DCGAN. The reason we choose PGAN rather than the more recent BigGAN (Brock et al., 2019) as the other compared method is that the two models achieved similar results without additional class information.

Comparison Results With CycleGAN

We compare the objective evaluation criteria (PSNR and SSIM) with CycleGAN[3] (Zhu et al., 2017) on the CelebA dataset. Note that SSIM and PSNR are not suitable for evaluating the quality of GANs, since GANs, generally, generate images directly from Gaussian white noise. That is, we do not have real images corresponding to the generated images, but real images are needed to calculate the PSNR and SSIM scores.

The reason we choose CycleGAN as the compared method is that it can be seen as a special kind of autoencoder model and, hence, can be used to calculate the PSNR and SSIM scores. The structure of CycleGAN is shown in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. As in the experiment of GFRSNs, we choose 65,536 training images and 16,384 testing images. For the training process, we divide the training set into two subsets of the same size, namely, A and B, to meet the unique circular training process. By training CycleGAN through 32,768 images in domain A and 32,768 images in domain B, we can calculate Train PSNR and Train SSIM. For the testing process, we also divide the testing set into two subsets of the same size, namely, A and B, to meet the unique circular training process. By training CycleGAN through 8,192 images in domain A and 8,192 images in domain B, we can calculate Test PSNR and Test SSIM. It can be known from the experimental process that in order to calculate the PSNR and SSIM values of the training data set and the testing data set, there are several characteristics when using CycleGAN:

(1) The training and testing processes are performed separately; that is, the trained generator of CycleGAN is not used in the testing process, since CycleGAN performs the task of image-to-image translation or style transfer (Gatys et al., 2016). In

[2] https://github.com/facebookresearch/pytorch_GAN_zoo

[3] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

TABLE 6 | Quantitative results of CycleGAN and the proposed GFRSNs with fractional orders $\alpha_1 = 0.4$, $\alpha_2 = 1$ on the CelebA dataset.

	Train PSNR	TestPSNR	Train SSIM	Test SSIM
Cycle GAN	30.8059	32.6890	0.9824	0.9822
Ours	27.9721	21.4631	0.9629	0.9350

order to get the Test PSNR and Test SSIM of the testing images, we still need to train CycleGAN with the testing images. For example, as is shown in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, the generator G_{AB} takes an image from domain A, and then tries to do an image-to-image translation, so that the output will be a fake image with a style similar to domain B. However, there is only one style of images in the CelebA dataset; therefore, the generator will learn the same image as the input. That is, it is unfair to use the PSNR or SSIM score to measure the quality of CycleGAN to some extent, since CycleGAN trains the testing images.

(2) In CycleGAN, the role of the generator is not focused on generating images from noise. On the contrary, the generator takes their effort to the task of image-to-image translation. When the style of two subsets is the same, this kind of image-to-image method will undoubtedly lead to pixel-level alignment and, hence, failure of pixel error-based metrics, such as PSNR and SSIM. That is, the PSNR and SSIM scores can be seen as the upper bound of other methods.

The results of the comparison of PSNR and SSIM scores of the proposed GFRSNs with CycleGAN are shown in **Table 6**, from which we can see that the result of GFRSNs is worse than that of CycleGAN, especially on the testing set. This is not surprising, because CycleGAN implements style transfer between training data and testing data, while GFRSNs implements reconstruction from FrScatNet features to images. The PSNR and SSIM scores of CycleGAN can be seen as the upper bound of GFRSNs; that is, the proposed GFRSNs still have a lot of room for improvement.

CONCLUSIONS

This study proposes generative fractional scattering networks (GFRSNs), which use fractional wavelet scattering networks (FrScatNets) as encoder to obtain features (or FrScatNet embeddings) and deconvolutional neural networks as decoder to generate an image. Additionally, this study develops a new feature-map fusion (FMF) method to reduce the dimensionality of FrScatNet embeddings. The impact of image fusion is also discussed in this study. The experimental results on the CIFAR-10 and CelebA datasets show that the proposed GFRSNs can lead to better generated images than the original GSNs in the testing dataset. Compared with GANs, the proposed GFRSNs lack details of the generated image because of the essence of the autoencoder structure; however, the proposed GFRSNs have the following merits:

(1) They can obtain an image code (or a latent vector), which is very helpful for downstream tasks such as image classification.

(2) They use learning-free FrScatNets instead of CNNs in the encoder stage, which significantly reduces the parameters.

(3) They may have a potentially good performance in the differential privacy (DP) learning framework, since Tramer and Boneh (2021) show that ScatNet outperforms deep CNNs in differential private classifiers. We studied the image generation performance of GFRSNs under the framework of differential privacy learning. **Appendix A** gives some preliminary results.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.cs.toronto.edu/~kriz/cifar.html; http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

AUTHOR CONTRIBUTIONS

JW did conceptualization, methodology, writing—reviewing, and editing. XQ did validation and revised and edited the manuscript. JZ did writing—original draft preparation, software, and visualization. FW did software, validation, and data curation. YK and GY did validation and project administration. LS did formal analysis, writing—reviewing, and editing. HS did supervision and resources. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Natural Science Foundation of China under Grant Nos: 61876037, 62171125, 31800825, 61871117, 61871124, 61773117, and 61872079 and in part by INSERM under the grant called IAL and IRP.

ACKNOWLEDGMENTS

We thank the Big Data Computing Center of Southeast University for providing facility support on the numerical calculations in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.752752/full#supplementary-material

REFERENCES

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisc. Rev. Comput. Stat.* 2, 433–459. doi: 10.1002/wics.101

Angles, T., and Mallat, S. (2018). "Generative networks as inverse problems with scattering transforms," in *2018 International Conference on Learning Representations (ICLR)* (Vancouver, BC).

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks," in *2017 International Conference on Machine Learning (ICML)* (Sydney, NSW), 214–223.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Brock, A., Donahue, J., and Simonyan, K. (2019). "Large scale GAN training for high fidelity natural image synthesis," in *2019 International Conference on Learning Representations (ICLR)* (New Orleans, LA).

Bruna, J., and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1872–1886. doi: 10.1109/TPAMI.2012.230

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). "InfoGAN: interpretable representation learning by information maximizing generative adversarial nets," in *2016 Proceedings of the 30th International Conference on Neural Information Processing Systems, (NeurIPS)* (Barcelona), 29, 2180–2188.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image style transfer using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 2414–2423.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27, 2672–2680. doi: 10.3156/JSOFT.29.5_177_2

Gregor, K., Papamakarios, G., Besse, F., Buesing, L., and Weber, T. (2018). "Temporal difference variational auto-encoder," in *2018 International Conference on Learning Representations (ICLR)* (Vancouver, BC).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). "Improved training of wasserstein GANs," in *2017 Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)* (Long Beach, CA) 30, 5769–5779.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2017). "beta-VAE: learning basic visual concepts with a constrained variational framework," in *2017 International Conference on Learning Representations (ICLR)* (Toulon).

Hudson, D. A., and Zitnick, L. (2021). "Generative adversarial transformers," in *Proceedings of the 38th International Conference on Machine Learning (ICML)* (Virtual), 4487–4499.

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of The 32nd International Conference on Machine Learning (ICML)* (Lille) 1, 448–456.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). "Progressive growing of GANs for improved quality, stability, and variation," in *2018 International Conference on Learning Representations* (*ICLR*) (Vancouver, BC).

Karras, T., Laine, S., and Aila, T. (2020a). A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 1:1. doi: 10.1109/TPAMI.2020.2970919

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020b). "Analyzing and improving the image quality of StyleGAN," in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Virtual), 8110–8119.

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *2015 International Conference on Learning Representations (ICML)* (Lille).

Kingma, D. P., and Welling, M. (2014). "Auto-encoding variational Bayes," in *2014 International Conference on Learning Representations (ICLR)* (Banff).

Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images.

Liu, L., Wu, J., Li, D., Senhadji, L., and Shu, H. (2019). Fractional wavelet scattering network and applications. *IEEE Trans. Biomed. Eng.* 66, 553–563. doi: 10.1109/TBME.2018.2850356

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2016). Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* 23, 1882–1886. doi: 10.1109/LSP.2016.2618776

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). "Deep learning face attributes in the wild," i*n 2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago), 3730–3738.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Mallat, S. (2012). Group invariant scattering. *Commun. Pure Appl. Math.* 65, 1331–1398. doi: 10.1002/cpa.21413

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv*.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). "Spectral normalization for generative adversarial networks," in *2018 International Conference on Learning Representations (ICLR)* (Vancouver, BC).

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)* (Haifa), 807–814.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). "f -GAN: training generative neural samplers using variational divergence minimization," in *2016 Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)* (Barcelona) 29, 271–279.

Oord, A., van den, Vinyals, O., and kavukcuoglu, koray. (2017). "Neural discrete representation learning," in *2017 Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)* (Long Beach, CA) 30, 6306–6315.

Qi, G.-J. (2020). Loss-sensitive generative adversarial networks on lipschitz densities. *Int. J. Comput. Vis.* 128, 1118–1140. doi: 10.1007/s11263-019-01265-2

Radford, A., Metz, L., and Chintala, S. (2016). "Unsupervised representation learning with deep convolutional generative adversarial networks," in *2016 International Conference on Learning Representations (ICLR)* (San Juan).

Razavi, A., Oord, A., van den, and Vinyals, O. (2019). "Generating diverse high-fidelity images with VQ-VAE-2," in *2019 Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)* (Vancouver, BC) 32, 14837–14847.

Rezende, D., and Mohamed, S. (2015). "Variational inference with normalizing flows," in *2015 Proceedings of the 32nd International Conference on Machine Learning (ICML)* (Lille), 1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic backpropagation and approximate inference in deep generative models," in *2014 Proceedings of the 31st International Conference on Machine Learning (ICML)* (Beijing), 1278–1286.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., et al. (2017). Progressive neural networks. *arXiv*.

Salimans, T., Kingma, D., and Welling, M. (2015). "Markov Chain Monte Carlo and variational inference: bridging the gap." in *Proceedings of the 32nd International Conference on Machine Learning (ICML)* (Lille) 37, 1218–1226.

Shi, J., Chi, Y., and Zhang, N. (2010). Multichannel sampling and reconstruction of bandlimited signals in fractional fourier domain. *IEEE Signal Process. Lett.* 17, 909–912. doi: 10.1109/LSP.2010.2071383

Simonovsky, M., and Komodakis, N. (2018). "GraphVAE: towards generation of small graphs using variational autoencoders," in *27th International Conference on Artificial Neural Networks (ICANN)* (Rhodes), 412–422.

Sohn, K., Yan, X., and Lee, H. (2015). "Learning structured output representation using deep conditional generative models," in *2015 Proceedings of the 29th International Conference on Neural Information Processing Systems, (NeurIPS)* (Montreal, QC) 28, 3483–3491.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). "Ladder variational autoencoders," in *2016 30th Annual Conference on Neural Information Processing Systems Conference (NeurIPS)* (Barcelona) 29, 3738–3746.

Tramer, F., and Boneh, D. (2021). "Differentially private learning needs better features (or much more data)," in *2021 International Conference on Learning Representations (ICLR)* (Virtual).

Wang, Y., Li, J., Yi, L., Yao, F., and J., Q. (2003). "Image quality evaluation based on image weighted separating block peak signal to noise ratio," in *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing* (Nanjing) 2, 994–97.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proc.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., She, Q., and Ward, T. E. (2021). Generative adversarial networks in computer vision: a survey and taxonomy. *ACM Comput. Surveys* 54, 1–38. doi: 10.1145/3459992

Yang, B., Zhong, J., Li, Y., and Chen, Z. (2017). Multi-focus image fusion and super-resolution with convolutional neural network. *Int. J. Wavelets Multiresol. Inform. Proc.* 15:1750037. doi: 10.1142/S0219691317500370

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2017). Loss functions for image restorationwith neural networks. *IEEE Transac. Comput. Imaging* 3, 47–57. doi: 10.1109/TCI.2016.2644865

Zhu, J. Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision* (*ICCV*), 2242–2251.

Zou, D., and Lerman, G. (2019). "Encoding robust representation for graph generation," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest), 1–9.

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fnbot.2021.762702

Integrated Circuit Board Object Detection and Image Augmentation Fusion Model Based on YOLO

Szu-Yin Lin[1]* *and Hao-Yu Li*[2]

[1] Department of Computer Science and Information Engineering, National Ilan University, Yilan City, Taiwan, [2] Department of Information Management, Chung Yuan Christian University, Taoyuan City, Taiwan

Industry 4.0 has been a hot topic in recent years. The process of integrating Cyber-Physical Systems (CPS), Artificial Intelligence (AI), and Internet of Things (IoT) technology, will become the trend in future construction of smart factories. In the past, smart factories were developed around the concept of the Flexible Manufacturing System (FMS). Most parts of the quality management process still needed to be implemented by Automated Optical Inspection (AOI) methods which required human resources and time to perform second stage testing. Screening standards also resulted in the elimination of about 30% of the products. In this study, we sort and analyze several Region-based Convolutional Neural Network (R-CNN) and YOLO models that are currently more advanced and widely used, analyze the methods and development problems of the various models, and propose a suitable real-time image recognition model and architecture suitable for Integrated Circuit Board (ICB) in manufacturing process. The goal of the first stage of this study is to collect and use different types of ICBs as model training data sets, and establish a preliminary image recognition model that can classify and predict different types of ICBs based on different feature points. The second stage explores image augmentation fusion and optimization methods. The data augmentation method used in this study can reach an average accuracy of 96.53%. In the final stage, there is discussion of the applicability of the model to detect and recognize the ICB directionality in <1 s with a 98% accuracy rate to meet the real-time requirements of smart manufacturing. Accurate and instant object image recognition in the smart manufacturing process can save manpower required for testing, improve equipment effectiveness, and increase both the production capacity and the yield rate of the production line. The proposed model improves the overall manufacturing process.

Keywords: smart manufacturing, Internet of Things, deep learning, YOLO, object recognition

Edited by:
Shin-Jye Lee,
National Chiao Tung University, Taiwan

Reviewed by:
Hao-Hsiang Ku,
National Taiwan Ocean University, Taiwan
Xingmei Li,
North China Electric Power University, China

****Correspondence:***
Szu-Yin Lin
szuyin@niu.edu.tw

Received: *22 August 2021*
Accepted: *22 October 2021*
Published: *11 November 2021*

Citation:
Lin S-Y and Li H-Y (2021) Integrated Circuit Board Object Detection and Image Augmentation Fusion Model Based on YOLO.
Front. Neurorobot. 15:762702.
doi: 10.3389/fnbot.2021.762702

INTRODUCTION

Smart manufacturing is based on smart factories involving artificial intelligence (AI), the Internet of Things (IoT), big data, and other technical tools. Smart manufacturing is the general term referring to an advanced manufacturing process and a system capable of perceiving information intuitively, making decisions automatically, and executing manufacturing processes automatically (Wang et al., 2018). In addition, it reports the current status of each device through the process of mechanical automation. Statistics and summarizing data can help us understand the device's condition or estimate its usable period. Moreover, smart manufacturing combines machines and deep learning technology to improve product quality and reduce costs. Consequently, the machinery has attained better production efficiency and adaptive maintenance time within the effective period. Providing better or more flexible services to customers is part of smart manufacturing's pursuit of true intelligence. Smart manufacturing is the focus of recent Industry 4.0 topics related to research and development or industry promotion. However, there are several issues in the implementation of smart manufacturing. Before the topic of smart manufacturing was formally proposed, the core concept in the background of automated manufacturing was the flexible manufacturing system (FMS) (Kimemia and Gershwin, 1983; Bihi et al., 2018). FMS hoped to establish a flexible and automated manufacturing engineering system in response to all predictable or unpredictable changes in the industry. However, this goal can only be achieved with the assistance of other technologies or systems (Yadav and Jayswal, 2018). In a process related to quality management inspection, although automated optical inspection (AOI) is applied, the screening standards are too high, and approximately 30% of the products are eliminated (Mukhopadhyay et al., 2019; Kovrigin and Vasiliev, 2020; Diering and Kacprzak, 2021). Moreover, this method requires a massive workforce and time to perform inspection in the second stage. In addition, only through the operator's correct implementation of various standard inspection procedures can it guarantee accurate manufacturing quality management. Therefore, a large number of professional employees undergo long-term training, increasing the labor cost. Smart manufacturing should include automated perception at its core and find a way to attain automated intelligence ultimately. In the process, various technologies and methods, such as intelligent image recognition and intelligent data analysis, can help achieve automatic identification and prediction. Auxiliary decision-making can also be used to perform automated execution in the environment, though it will be challenging.

As AI image recognition becomes more and more mature nowadays, the combination of deep learning with classic computer vision has become a trend. Today, most mainstream technology for image recognition applications uses convolutional neural networks (CNNs). Since the re-emergence of deep learning in 2012, scholars and experts have proposed several new methods to solve the problems encountered by neural networks in the past. The shortcomings of CNNs in the past have also been reduced (Khan et al., 2020). In recent years, the characteristics of graphics processing units have also been fully utilized to accelerate the calculation of deep learning algorithms; therefore, the algorithm's efficiency has dramatically improved. The most crucial technological turning point in image recognition is the development of the region-based convolutional neural networks (R-CNN) algorithm. This technology first solved the problem of the insufficient dataset, and later, the related models introduced also performed well in terms of performance and recognition accuracy (Bharati and Pramanik, 2020). Based on it, the Faster R-CNN algorithm was developed, which allows the calculation speed of the algorithm to reach a different level of sophistication. As a result, image recognition technology is getting closer and closer to the goals of achieving both high speed and high precision (Gavrilescu et al., 2018; Maity et al., 2021).

Nowadays, several cases of the combination of computer vision with deep learning of the IoT have been implemented, and many positive feedbacks have been obtained in academic research and real-life applications (Wang et al., 2020; Xu et al., 2020; Lian et al., 2021). Accurate image recognition technology helps classify product types, confirm product integrity in an actual field, and helps establish a smart manufacturing field. The method proposed in this study is based on the R-CNN-related model of the deep learning method. The integrated circuit board (ICB) image is selected as the dataset to complete the image recognition model. The first stage aims to acquire different types of ICB images for model training. Thus, we first constructed the initial phase of image recognition so that the model can understand the characteristics of different types of ICBs and their details. In the second stage, a camera is used for real-time identification of the smart manufacturing field by collecting real-time images and returning the data to the server for data analysis, thereby solving the FMS's quality management inspection and monitoring. This study has three main objectives: (1) to establish an image recognition model that is suitable for use in the smart manufacturing field; (2) to explore the image augmentation fusion and optimization method of the model so that the model can learn more image features to improve the accuracy of image recognition; and (3) to solve the problem of over screening in automatic optical inspection and introduce the model into practical applications to test the directionality of ICB images.

LITERATURE REVIEWS

R-CNN and SPP-Net

There are three main problems to be solved by region-based convolutional neural networks (R-CNN), which involve (1) accuracy of object recognition; (2) whether more feature values can be obtained; and (3) solving the problem of insufficient dataset. Compared with previous CNNs, R-CNN proposes a method for selecting region proposals of selective search (Girshick et al., 2014) to increase its dataset and find critical features. Previously, when solving dataset problems, the data augmentation method mentioned in "ImageNet Classification with Deep Convolutional Neural Networks" was first considered (Krizhevsky et al., 2012). Notably, the R-CNN region proposal's concept also aims at this problem (Girshick et al., 2014). In R-CNN, the input of selective search (Girshick et al., 2014) is an image, and the output is the possible position of the object. The principle is to initialize a similar empty set in advance, calculate the similarity of all adjacent intervals, store it in the empty set, find the region with the highest degree of similarity, and return it to the final total set. The region in the total set is the object's bounding box, and the similarity is judged based on color, texture, size, and shape, and iteratively combining similar regions to form objects. R-CNN obtains many region proposal images through the selective search method, but still needs to use the same image size as the input of the entire neural network because the fully connective layer in CNN must maintain the exact dimensions in operation, and the operation parameters also need to consider the upper-layer relationship. However, spatial pyramid pooling network (SPP-Net) addressed this issue: by adding a layer of SPP before building the fully connective layer. The function and principle of SPP are that the data process is performed before regular data input to comply with the fully connective layer problems mentioned above. SPP replaces the last pooling layer before the fully connective layer, and to adapt to the feature maps of different resolutions, the layer is defined as a scalable pooling layer so that a fixed ratio can be used through SPP. This way of converting and maintaining the input of the fully connective layer is a breakthrough in this part (He et al., 2014).

Fast R-CNN and Faster R-CNN

The core problem with R-CNN is that it generates a large number of region proposal images through selective search. When pre-processing data, we still have to refer to the data augmentation method by AlexNet (Krizhevsky et al., 2012) to make modifications, which may lead to the loss of features of the original region proposal. In addition, if each region proposal image is put into training, the vast computational waste caused by repeated feature extraction will make the model inefficient. With SPP-Net, it is still time-consuming to train all the images, so Fast R-CNN was created. Fast R-CNN is a unified version of R-CNN and SPP-Net. Fast R-CNN proposed RollPooling (region of interest pooling), which uses the idea of SPP-Net to do the conversion work into the fully connective layer based on the input image. First, the original image is convolved to generate a feature map corresponding to RollPooling. Then the image trained in the region proposals is directly given the convolution value of the region proposal image through RollPooling to do MaxPooling. The most significant advantage of RollPooling is the increase of massive processing speed. Besides, regardless of the size of the given feature map, the dimensions of the output data can be kept uniform (Girshick, 2015). The key problem Fast R-CNN (Girshick, 2015) wants to solve is calculating the image of the region proposals. Hence, an improved Faster R-CNN was developed to solve the issue of repetitive region proposals directly. It does not abandon the selective search (Girshick et al., 2014) method but finds region proposals with features more efficiently. Therefore, the concept of RPN (region proposal network) is proposed in the Faster R-CNN architecture. The core concept of RPN is not to find the region proposals from the original image but to find the region proposals through the convolved feature map of the original image as input. The RPN extracts region proposals through a sliding window, and each sliding window generates nine different size of windows (anchor box). After removing the corresponding nine window features, the extra part is discarded, and the anchor box with an overlap area value >0.7 as the foreground is calculated. The overlapping area is set to the background, the most suitable region proposals feature map is found, and the concept of RollPooling is combined to train the model. This method is very similar to Fast R-CNN in terms of results and has dramatically improved the speed. It is also one of the most commonly used models in R-CNN (Ren et al., 2016).

YOLO

After introducing Faster-RCNN (Ren et al., 2016), You Only Look Once (YOLO) (Redmon et al., 2016) and ordinary R-CNN were introduced in the same year with different architectures. The past versions of R-CNN, from selective search (Girshick et al., 2014) to RPN, were all intended to increase training and reduce energy consumption. Although the development of RPN enables sharing of convolution values, YOLO uses an end-to-end method for object detection using an entire image as the input of the neural network to predict the coordinate position of the bounding box directly. YOLOv1 is fast in calculation and can be applied to real-time fields, but the prediction of the position is not accurate enough, and the performance of small object fields is poor. In addition, for object images' recognition, it is impossible to distinguish between the foreground and background of the object effectively. Interestingly, YOLOv2 (Redmon and Farhadi, 2016) imported the anchor box to increase accuracy. The original YOLOv1 version divides the image into 7 × 7 grids, and each grid predicts two bounding boxes, which is better than importing 1,000 pre-selected regions into the anchor box. The fully connective layer was removed and changed to a fully convolutional network, and dropout was removed to optimize the overall speed and accuracy of YOLOv3 (Redmon and Farhadi, 2018). The maximum input of the image can reach 608 × 608 pixels, and many optimizations have been made. For example, residual neural network (ResNet) and feature pyramid network (FPN) are used to improve the detection of small objects; the darknet53 network is applied; the detection threshold of YOLO model can be adjusted in the training process according to the threshold parameter in its network architecture. Faster R-CNN's architecture RetinaNet is built using ResNet. Comparing YOLOv3 with ResNet, it can be observed that YOLOv3 can achieve the same results in a relatively short time. The mentioned FPN architecture uses three boxes of different sizes. The model can learn the image characteristics of different blocks through these three scales to improve YOLO's shortcomings in small object prediction (Redmon and Farhadi, 2018). YOLOv4 (Bochkovskiy et al., 2020) has improved the previous version in many aspects. The author uses the Mosaic method, which used random scaling and cropping to mix and stitch 4 kinds of pictures from the original datasets, to enrich the data set and enhance the stability of the model for small target detection. For stability, the network uses CSPDarknet53, which is composed of darknet53 and CSPNet (Wang et al., 2019), which greatly reduces equipment requirements and computing speed. The author also drew on the PANet (Path Aggregation Network) (Liu et al., 2018) used in the field of image segmentation, integrates PAN on the basis of the FPN architecture, and adds SPP (Spatial pyramid pooling) to improve the ability of feature extraction.

MATERIALS AND METHODS

Nowadays, in implementing smart manufacturing, intelligence should be implemented to achieve the most effective results to

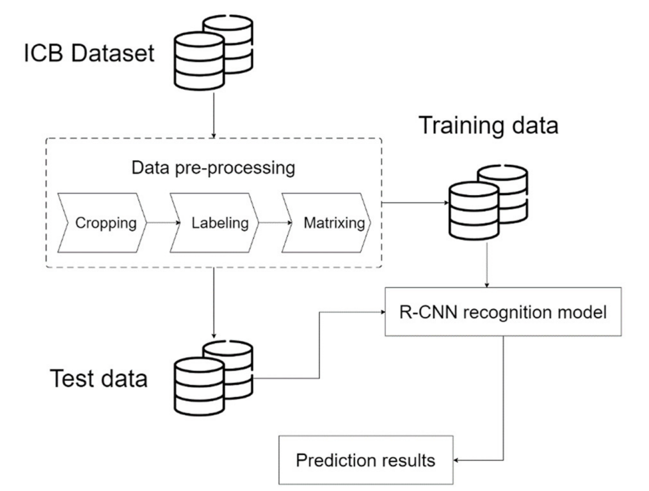


FIGURE 1 | Image recognition and object detection model for ICB.

complete the quality management part of FMS effectively. In the field of traditional non-intelligent manufacturing, several problems are encountered. (1) Although the current automatic optical inspection method can achieve accurate inspection, its parameter setting is too strict, resulting in a pass rate of ∼70%. It is still necessary to employ field operators to complete the second inspection stage to ensure the yield. (2) In traditional manual monitoring, the biggest problem is that people may suffer from mistakes due to inattention or fatigue, which affects the quality of some parts. (3) In the field of smart manufacturing, the inspection process should give high accuracy in real-time. Therefore, we must find a suitable image recognition model to apply here. The deep learning image recognition method allows the selected model to learn the item's features by using the features provided in the dataset. Consequently, accurate image recognition in the manufacturing system can be attained, and the integrity and quality inspection of ICBs can be completed through precise image recognition. This study aims to build an image recognition model of ICBs so that various types of ICBs can be classified in this model according to the system architecture flow of this study, such as **Figure 1**. In the image recognition and object detection model for ICB, the first stage is to collect part of the dataset and establish the image database standard that can be used based on the R-CNN method. Then, the collected images are cropped, feature labeled, and matrixed. Later, the dataset is divided into training data and test data. Next, an R-CNN is constructed to train the image recognition model. Finally, the image of the test data is mapped to the recognition model to generate the result. The results are respectively sent to the user and the server end for data analysis applications.

ICB Data Collection

The training data in this study has five types of ICB images, and 100 images are collected based on these five types. The ICB images used in this study must contain identifiable features under specified conditions. First, training the model for collecting images is standardized to better sample the image features in the data collection part. While collecting images, two methods of data collection can be used. In both scenarios, the ICB that needs to be pictured must be placed in the center of the image and then divided into near and far for feature collection. Moreover, in the collection process, the background is changed to be used for image recognition under different backgrounds. The focus of long-range shooting has covered the entire ICB. On the contrary, the focus of short-range shooting is mainly on the integrity and clarity of the internal structure of the ICB. Both methods must sample the different angle characteristics of the ICB during the shooting process. At least 100 samples of each category must be tested, and the final data collection shall be based on the five types of ICBs.

Pre-process

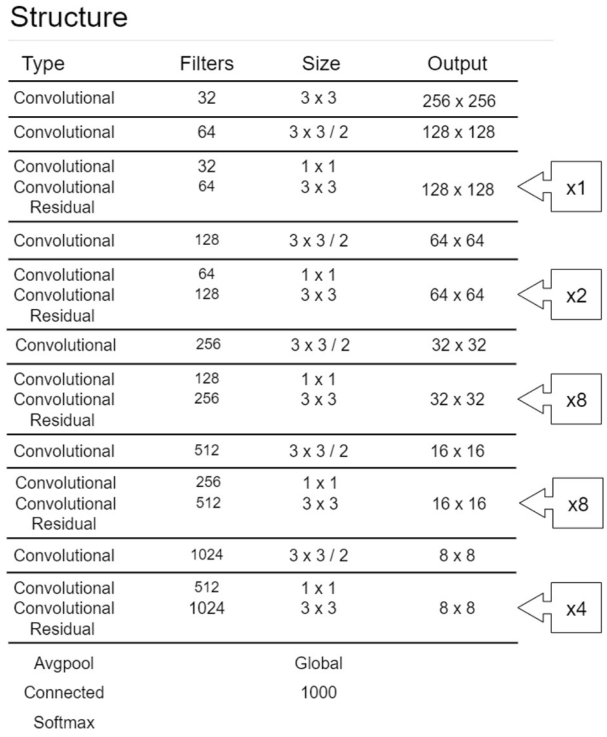
To successfully import the dataset into the model's training process, pre-processing must be done. The purpose of data pre-processing is to keep the input data in a consistent form, such as fixed image size or labeling so that it fits within the processing range of the R-CNN model before entering the model training process. The pre-processing of the data here includes three steps: the first step is to cut each ICB dataset into the size of 1,024 × 1,024 pixels without losing key details of the board. Only then can the dataset be easily imported into the model. The second step is to mark the image area through the open-source software Labellmg. Labellmg is the most commonly used software for labeling images. For our classification, we can mark the features in the image by selecting the box. The third step is to carry out matrix work. The image recognition model is different from humans. Humans capture features through images viewed by their eyes. Machines, on the other hand, use a data matrix to understand the key features in blocks in two-dimensional images and then use this matrix in the model for the application.

To train the YOLO model more effectively, pre-processing must be carried out for the first stage of data collection. The purpose is to make the model more focused on learning features with organization and clarity when learning images. In this stage, we must first set a fixed image size to mark the learning features of the model and then, convert the marked features into a matrix to train the neural network model. The steps are as follows:

- Image cutting: Use ImageSplitter, an online image cutting tool on the Internet, to fix the image size to capture the characteristics of each image and define the fixed size as 1,024 × 1,024 pixels.
- Data label matrix: Use the open-source software Labellmg to label images and feature matrix for training the model to correspond to the features that this study hopes to learn to complete the full model training.

Model Selection

There have been many studies comparing model suitability for smart manufacturing. In this study, YOLO is selected as the model. In the past, when recognizing R-CNN in images, most of them used the model architecture of Faster-RCNN for image recognition. Indeed, the accuracy of Faster-RCNN is still the



Structure

Type	Filters	Size	Output	
Convolutional	32	3 x 3	256 x 256	
Convolutional	64	3 x 3 / 2	128 x 128	
Convolutional Convolutional Residual	32 64	1 x 1 3 x 3	128 x 128	x1
Convolutional	128	3 x 3 / 2	64 x 64	
Convolutional Convolutional Residual	64 128	1 x 1 3 x 3	64 x 64	x2
Convolutional	256	3 x 3 / 2	32 x 32	
Convolutional Convolutional Residual	128 256	1 x 1 3 x 3	32 x 32	x8
Convolutional	512	3 x 3 / 2	16 x 16	
Convolutional Convolutional Residual	256 512	1 x 1 3 x 3	16 x 16	x8
Convolutional	1024	3 x 3 / 2	8 x 8	
Convolutional Convolutional Residual	512 1024	1 x 1 3 x 3	8 x 8	x4
Avgpool		Global		
Connected		1000		
Softmax				

FIGURE 2 | YOLOv3_Darknet53 network (Redmon and Farhadi, 2018).

TABLE 1 | Comparison table based on YOLO model.

Model	Advantages	Disadvantages
YOLOv3	Benchmark	Benchmark
YOLOv3_tiny	Fast training and lightweight architecture	The number of model layers is low, and it is difficult to reach the maximum value
YOLOv3_voc	Low confidence threshold and small input image	Features are relatively easy to lose focus
YOLOv3_spp	Can be used with the multi-scale conversion of eigenvalues	Features are easily compressed during conversion

highest, but to deal with the field of smart manufacturing, real-time recognition of images is vital. YOLO has a faster real-time response speed with an accuracy of results close to Faster-RCNN. Therefore, this study uses YOLO as the R-CNN model of the architecture. **Figure 2** shows YOLO's network configuration diagram (Redmon and Farhadi, 2018). YOLO is a multi-level R-CNN, where the first layer defines the dimensions of the input parameters and the output layer performs classification actions according to its final output results. Thus, the hidden layer between the input and output layers is the main structure of this R-CNN. The activation function used after each CNN layer is Leaky ReLU, and Residual refers to the ResNet architecture, which replaces the activation function covering the two-layer CNN. The functions and tasks of each layer are as follows:

- Input layer: After an ICB image is cut into the input size of the model, the learning features are marked. Then the parts are converted into a matrix pattern that the machine can understand, thereby becoming the model's input data.
- Convolution layer: The ICB image is two-dimensional in this study, so a two-dimensional convolutional layer is used. The convolutional layer can parameterize the image of the ICB through the size of its image, the kernel size, and the feature factor.
- Leaky ReLU layer: This derivation of ReLU uses the function in the neural network node to increase the non-linear characteristics of the entire neural network function and define the node's output so that it is suitable for solving the dying ReLU problem.
- Residual layer: Its original name is the residual network (ResNet) and its core is residual block. To solve the problem of an unexpected increase in the error rate during training, some of the weight parameters may tend to zero or become zero during the regular conversion of each layer, and the error rate will increase when the best solution is ignored.
- Average pooling layer: This layer replaces the fully connective layer used at the end of the general neural network. The most significant disadvantage of the fully connective layer is that the number of parameters is too large, resulting in overfitting. Therefore, the average pooling layer replaces the weighted connection layer to directly give each feature its sense to prevent the overfitting problem caused by the fully connected layer.
- Softmax layer: The Softmax layer multiplies the weight matrix, adds the characteristic error to generate the Softmax function, and applies it to the output of the average pooling layer.
- Classification layer: The classification layer obtains the output of the previous Softmax layer and classifies the input data according to the final output.

This study is built on four models based on YOLOv3, namely, YOLOv3, YOLOv3_tiny, YOLOv3_voc, and YOLOv3_spp. The comparison of these four models is shown in **Table 1**.

- YOLOv3: It is the third version of the initial model of YOLO, which adds the model architecture of Darknet53 and a multi-scale method to verify the feature map. The multi-scale approach helps the model learn the detailed features of the image through three different sizes, which is a breakthrough for YOLOv3. In addition, it can use images up to 608 × 608 as input data (Redmon and Farhadi, 2018).
- YOLOv3-tiny: There are 19 layers of CNN, which is a part of the gap compared with the 75 layers of the original version. Its advantage is that it has better applications for devices with limited computing resources and fast training.
- YOLOv3-voc: It is an improvement of YOLOv3. The original input of YOLOv3 is 608 × 608, and YOLOv3-voc is 416 × 416, which is the same as that of YOLOv3-tiny. This method focuses on retaining the convolutional layer, reducing the image size to improve the training speed, and reducing its

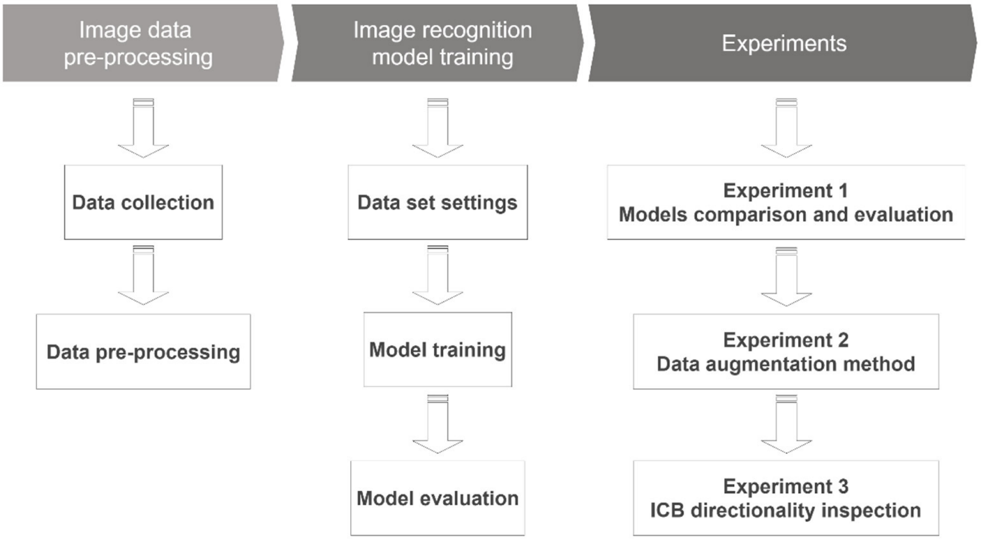


Image data pre-processing
Image recognition model training
Experiments
Data collection
Data pre-processing
Data set settings
Model training
Model evaluation
Experiment 1 Models comparison and evaluation
Experiment 2 Data augmentation method
Experiment 3 ICB directionality inspection

FIGURE 3 | Research flow chart.

ignore thresh (the threshold value that the overlapped block of the predicted labeled area and the overlapped labeled area must exceed) for training.

- YOLOv3-spp: The purpose of adding the SPP to YOLOv 3 is to convert the selected feature maps to the same size using the SPP fixed-scale conversion method to achieve a more accurate learning feature model training (Huang and Wang, 2019).

Model Adjustment

YOLO's overall training process includes classification design, training dataset cutting, test dataset cutting, naming of each category, and parameter settings in order. These five items are YOLO's current framework, and the selection and setting of the datasets and models are used to complete the image recognition work. In this process, the related settings of model adjustment are introduced as follows: Classes: identify target types; Train: training dataset settings; Valid: verify dataset settings; Names: specify the name of the target type; and Backup: store model parameters. During the model training process, YOLO trains the recognition model based on the training data. After repeated iterative training, the image recognition and object detection results are generated according to the model parameters and the classification settings. This result has the characteristics of the relevant image data in the learning process. Finally, the membership classification is marked when an output is achieved, and the overall recognition accuracy is returned. The parameter setting values when using the learning model in this study are as follows. (1) Batch: 16 (refers to the number of batches that have passed to update the parameters once); (2) Subdivisions: 4 (if the memory is insufficient, the batch will be divided into sub-batches); (3) Width: 608 (the width of the input image data); (4) Height: 608 (the height of the input image data); and (5) Momentum: 0.9 (in neural networks, it is a variant of the stochastic gradient descent. It replaces the gradient with a momentum, which is an aggregate of gradients); (6) Decay: 0.0005 (parameter weight attenuation setting to prevent overfitting); and (7) Learning rate: 0.001 (initial learning rate). The study process includes data collection, pre-processing methods, experimental environment, model establishment, discussion, evaluation, and analysis to verify the proposed R-CNN image recognition model design method applied to the smart manufacturing field. **Figure 3** shows the flow chart of the study.

EXPERIMENTS

Evaluation Metrics

(1) Mean Average Precision (mAP): As shown in equation (2), the accuracy of all classifications is averaged (an average is calculated by estimating the prediction and actual accuracy). The basic accuracy calculated is as follows:

TP(ICB1), True Positive in ICB1: The classification result of the current model is correct, and the overlap between the predicted labeled area and the actual labeled area is high enough.

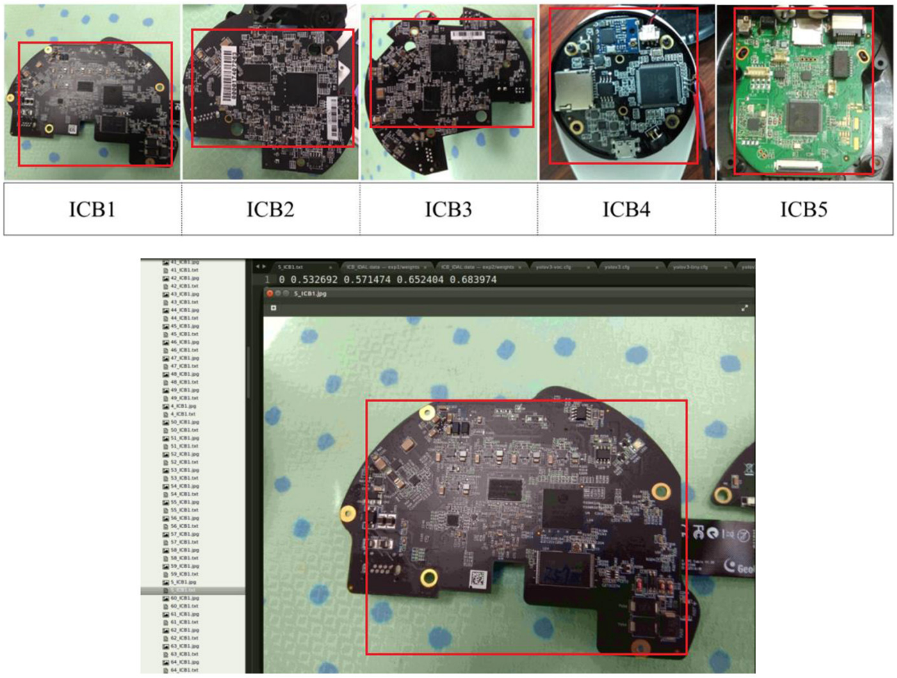
FP(ICB1), False Positive in ICB1: The classification result of the current model is incorrect, or the overlap between the predicted labeled area and the actual labeled area is not high enough.

From this, the accuracy of classification ICB1 can be calculated from the following equation (1):

$$\boldsymbol{Precision\ (ICB1)} = \frac{\boldsymbol{TP(ICB1)}}{\boldsymbol{TP\ (ICB1) + FP(ICB1)}} \tag{1}$$

Therefore, the mAP of each category is calculated from equation (2) (take N categories as an example):

$$\boldsymbol{mAP} = \frac{\boldsymbol{Precision\ (ICB1) + \ldots + Precision(ICBN)}}{\boldsymbol{N}} \tag{2}$$



```
1  0 0.532692 0.571474 0.652404 0.683974
```

FIGURE 4 | ICB type table and image labeling.

(2) Recall: The ratio of the number of correctly identified categories in the prediction result to the target in the test data, calculated from equation (3).

$$\textbf{\textit{Recall}}\,(\textbf{\textit{ICB1}}) = \frac{\textbf{\textit{TP}}(\textbf{\textit{ICB1}})}{\textbf{\textit{TP}}\,(\textbf{\textit{ICB1}})\,+\textbf{\textit{FN}}(\textbf{\textit{ICB1}})} \tag{3}$$

TP(ICB1), True Positive in ICB1: The classification result of the current model is correct, and the overlap between the predicted labeled area and the actual labeled area is high enough.

FN(ICB1), False Negative in ICB1: It means that the current model test set is not classified in the pre-set classification, and the recognition model classifies it as one of the classifications.

Experimental Designs

This study uses the evaluation indicators of the YOLO image recognition model to compare the image recognition results of four different models of YOLO and enhance the difference in the size of the training data through the image augmentation fusion method. The following three aspects are used to evaluate the performance of the proposed method.

- Models comparison and evaluation: This study identifies four different models based on YOLOv3 and use fixed parameters to train the model. In addition, five different types of ICB images are used; each type has 100 images, with 80 of them used for training and 20 for verification. Thus, the total dataset contains 400 training images and 100 verification images. Finally, an additional 60 images are used as a test.
- Data augmentation: In this stage, each classification's original ICB images are used for data augmentation methods. The amplification parameters used are rotation_range, width_shift_range, height_shift_range, shear_range, zoom_range, horizontal_flip, vertical_flip, and fill_mode. The 100 original images of each classification are processed by the data augmentation method to generate 500 images, and then 400 images per classification are used as the model's training data. The remaining 100 images are used as verification data. There are a total of 1,600 training images and 400 verification images. Finally, the same 60 test data are used to discuss the analysis of the data augmentation method for the model feature training and learning.
- ICB directionality inspection: This stage of the experiment checks the core image of the integrated circuit board to see whether the chip is installed incorrectly. Type 5 of the ICBs is used to perform this test. The whole experiment uses 88 training images 22 verification images, and 50 test images. These images contain both correct and incorrect integrated circuit images (incorrect images are ICBs with wrong core directionality). The images are inspected to see whether the model can correctly check the core installation error of the ICB. This experimental model uses the best model discussed in the 1st and 2nd experiments for training.

Training Dataset

As shown in **Figure 4**, this study collects 100 images of each of the five types of ICB, and the data must be labeled during YOLO training. After labeling each image, the image is set to the learning format of the YOLO model on the Darknet platform corresponding to its classification. Images of each format are

TABLE 2 | Experiment 1-various YOLO models result comparison table.

Model	Average iteration time	Average error	Average training accuracy	Training recall rate	Average test accuracy	Test recall rate	Maximum average accuracy
YOLOv3	0.94	0.05	98.87%	98%	91.39%	88%	91.39%
YOLOv3-tiny	0.18	0.203	98.82%	99%	91.63%	87%	91.73%
YOLOv3-voc	0.72	0.036	99%	100%	91.66%	90%	91.9%
YOLOv3-spp	0.98	0.057	98.47%	99%	87.86%	87%	88.68%

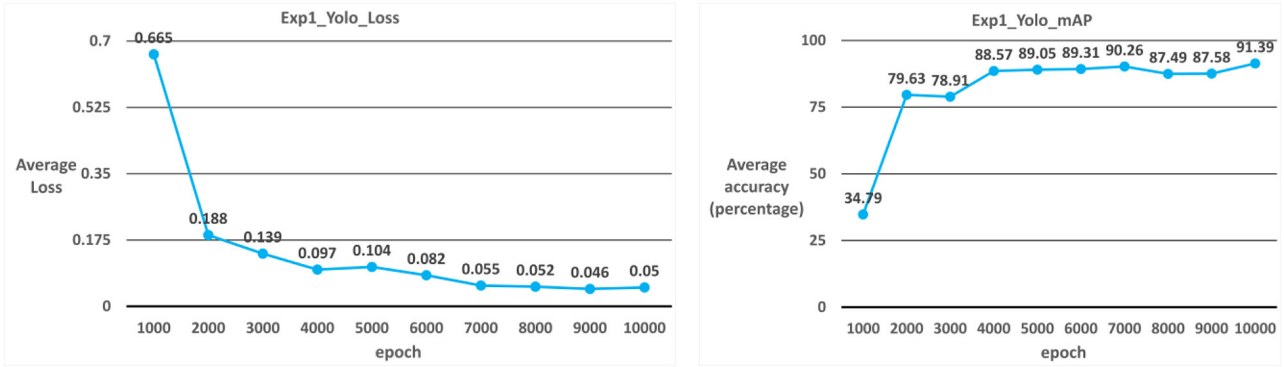


FIGURE 5 | YOLOv3 average error rate and average accuracy rate.

divided into training and verification data to complete the preliminary model training settings.

Experimental Results

Models Comparison and Evaluation

During the model training process, we use the YOLOv3 model with the parameters that have been set, and the training iteration target is 10,000. During the training, the values are stored as train_log_loss.txt file to help us understand each iteration's error value and average error, the current learning rate, the number of training images, and the training time. The entire training set includes 500 ICB images, which are classified into five categories, of which 100 ICB images are used as the training phase verification of the overall model, and the number of training iterations is 10,000. Then, using the trained model parameters, the current classification status of each classification and the generation of mAP and recall of the model are calculated through the additional 60 images of test data. In this stage, the four models YOLOv3, YOLOv3-tiny, YOLOv3-voc, and YOLOv3-spp are presented in sequence from Case 1 to Case 4, respectively, showing the training process and the accuracy during training and the final test accuracy. Experimental discussion in **Table 2** shows that the YOLOv3-voc model is significantly better than the other three in 10,000 iterations. Experiment 1 shows that the YOLOv3-voc model is the best, and its overall average error is 0.036, and its maximum average accuracy is 91.9%.

Case 1: YOLOv3

The model used in Case 1 is the YOLOv3 model. As shown in **Table 2**; **Figure 5**, the average iteration time is 0.94 s, and the average error rate is 0.05. Therefore, the average accuracy rate in the training phase can reach 98.87%, and the recall rate can reach 98%. During the test phase, 60 ICB images are used as test data. As a result, the average accuracy rate in the test phase can reach 91.39%, and the recall rate can reach 88% due to the overall model performance.

Case 2: YOLOv3-tiny

The model used in Case 2 is the YOLOv3-tiny model. As shown in **Table 2**; **Figure 6**, the average iteration time is 0.18 s, and the average error rate is 0.203. Therefore, the average accuracy rate in the training phase can reach 98.82%, and the recall rate can reach 99%. During the test phase, 60 ICB images are used as test data. As a result, the average accuracy rate in the test phase can reach 91.63%, and the recall rate can reach 87% due to the overall model performance.

Case 3: YOLOv3-voc

The model used in Case 3 is the YOLOv3-voc model. As shown in **Table 2**; **Figure 7**, the average iteration time is 0.72 s, and the average error rate is 0.036. Therefore, the average accuracy rate in the training phase can reach 99%, and the recall rate can reach 100%. During the test phase, 60 ICB images are used as test data. As a result, the average accuracy rate in the test phase can reach 91.66%, and the recall rate can reach 90% due to the overall model performance.

Case 4: YOLOv3-spp

The model used in Case 4 is the YOLOv3-spp model. As shown in **Table 2**; **Figure 8**, the average iteration time is 0.98 s, and the

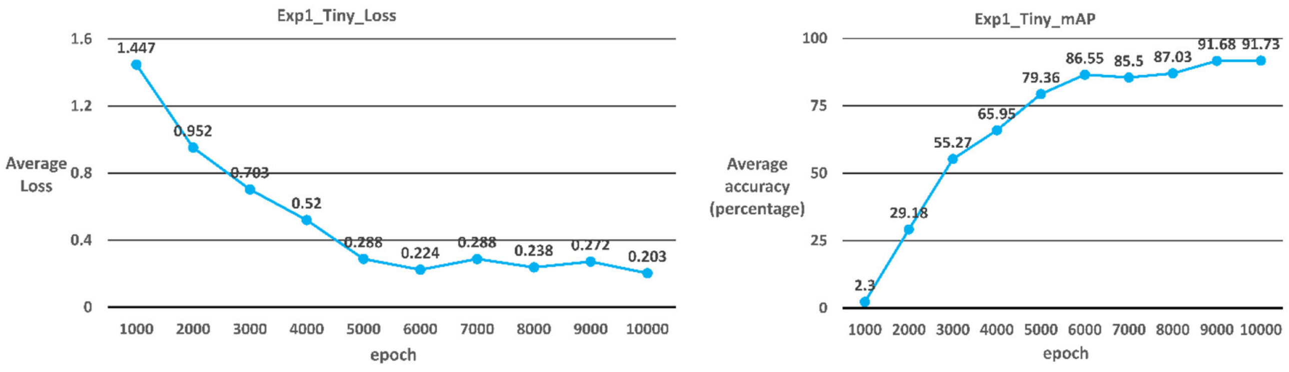


FIGURE 6 | YOLOv3_tiny average error rate and average accuracy rate.

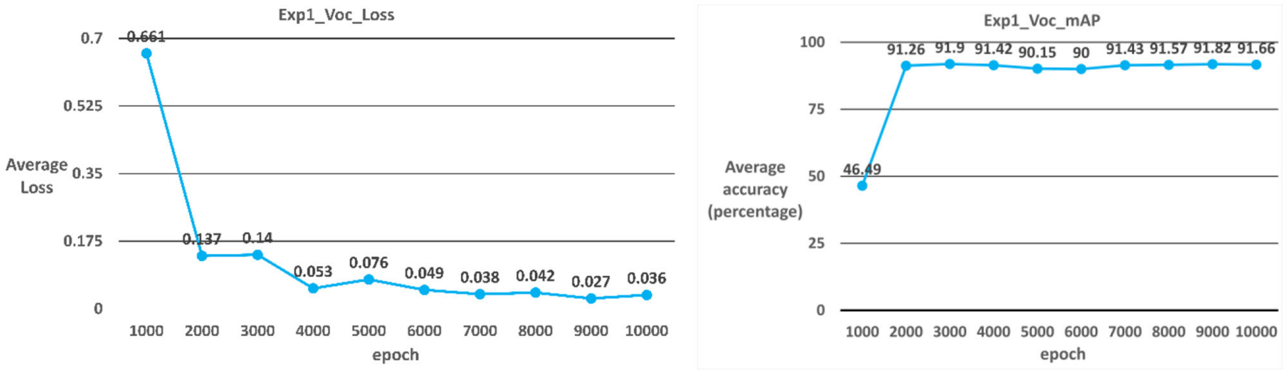


FIGURE 7 | YOLOv3_voc average error rate and average accuracy rate.

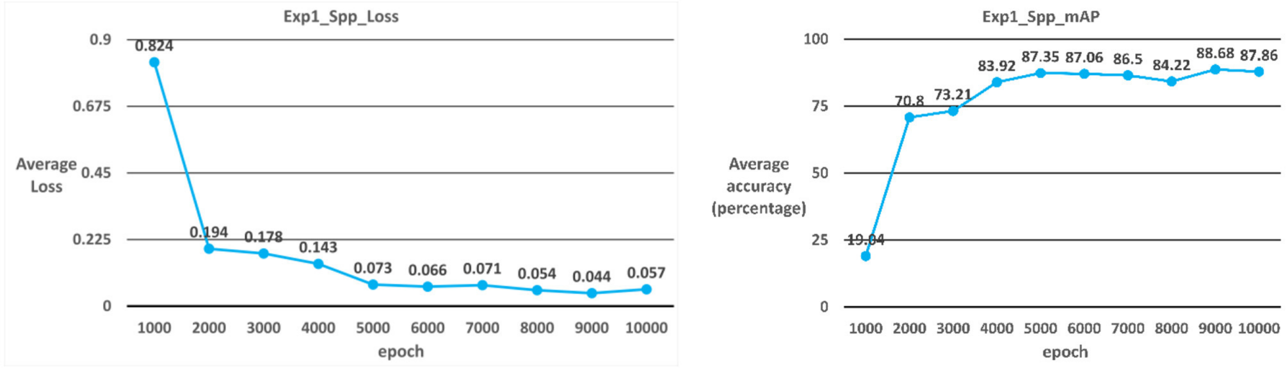


FIGURE 8 | YOLOv3_spp average error rate and average accuracy rate.

average error rate is 0.057. Therefore, the average accuracy rate in the training phase can reach 98.47%, and the recall rate can reach 99%. During the test phase, 60 ICB images are used as test data. As a result, the average accuracy rate in the test phase can reach 87.86%, and the recall rate can reach 87% due to the overall model performance.

Data Augmentation

In this stage of the experiment, the impact of the amount of data on training is discussed in advance, so data augmentation methods are used to increase the dataset. The result of a single image using the data augmentation method is shown in **Figure 9**, and the image generated by the data

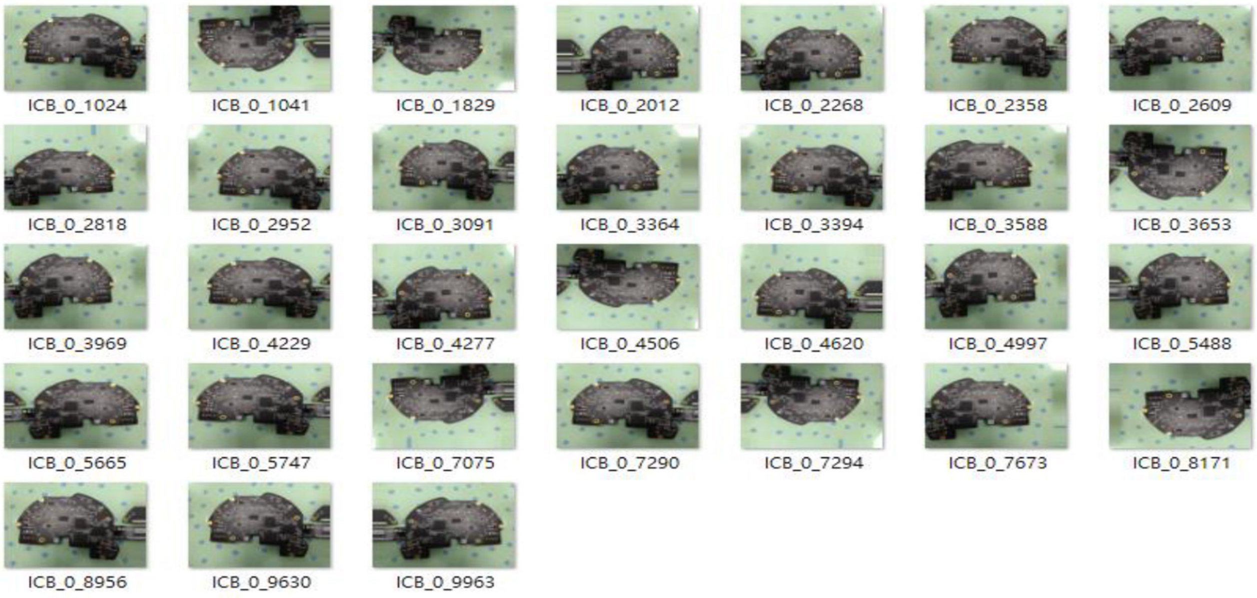


FIGURE 9 | Results of data augmentation methods.

TABLE 3 | Experiment 2-various YOLO models result comparison table (after data augmentation).

Model	Average iteration time	Average error	Average training accuracy	Training recall rate	Average test accuracy	Test recall rate	Maximum average accuracy
YOLOv3	0.67	0.105	99.62%	99%	91.14%	92%	94.86%
YOLOv3-tiny	0.20	0.251	99.55%	97%	93.56%	90%	94.87%
YOLOv3-voc	0.71	0.06	99.8%	100%	94.72%	95%	96.53%
YOLOv3-spp	0.66	0.079	97.07%	97%	92.22%	95%	94.58%

augmentation method still requires data pre-processing and labeling.

In this stage of the experiment, as shown in **Table 3**, the performance of the YOLOv3-voc model at the number of iterations of 10,000 is significantly better than the other three. The experimental result of experiment 2 is that the YOLOv3-voc model is the best. It has an average error value of 0.06, and the highest average accuracy rate can reach 96.53%.

Comparing the results from the YOLOv3-voc model of experiment 1 and experiment 2, listed in **Tables 2**, **3**, respectively, it is found that using data augmentation methods to allow the model to learn more image features can significantly improve its average accuracy and recall rate.

ICB Directionality Inspection

A total of 160 images of type-5 integrated circuit board model (ICB5) are used in this experiment stage. In the experiment, the images are divided into 88 for training datasets, 22 for verification datasets, and 50 for test datasets; all datasets contain both correct and incorrect integrated circuit images. The model used is the YOLOv3_voc model, and the model is trained to 10,000 iterations. The identification results are shown in **Figure 10**, showing the correct identification and three kinds of incorrect identification. Correct: The direction of the ICB recognition image is correct; Error type 1: The direction of the ICB recognition image shows type one error; Error type 2: The direction of the ICB recognition image shows type two error; Error type 3: The ICB recognition image direction shows type three error; None: Cannot identify the direction of the ICB identification image. For the result, among the 50 test images, only one image is currently not recognized. The original training model and actual prediction results are shown in **Table 4**, showing a correct rate of 98%, which is more than 90% required for general applications. Furthermore, the recognition time for each image is no more than one s, which is practical for smart manufacturing fields that require real-time recognition.

SUMMARY

The experiment in this study is divided into four stages. In the first stage, we must execute the pre-processing of the dataset to complete the learning goal and then generate a complete training process. The second stage focuses on the four models under YOLOv3 to explore more suitable model for smart

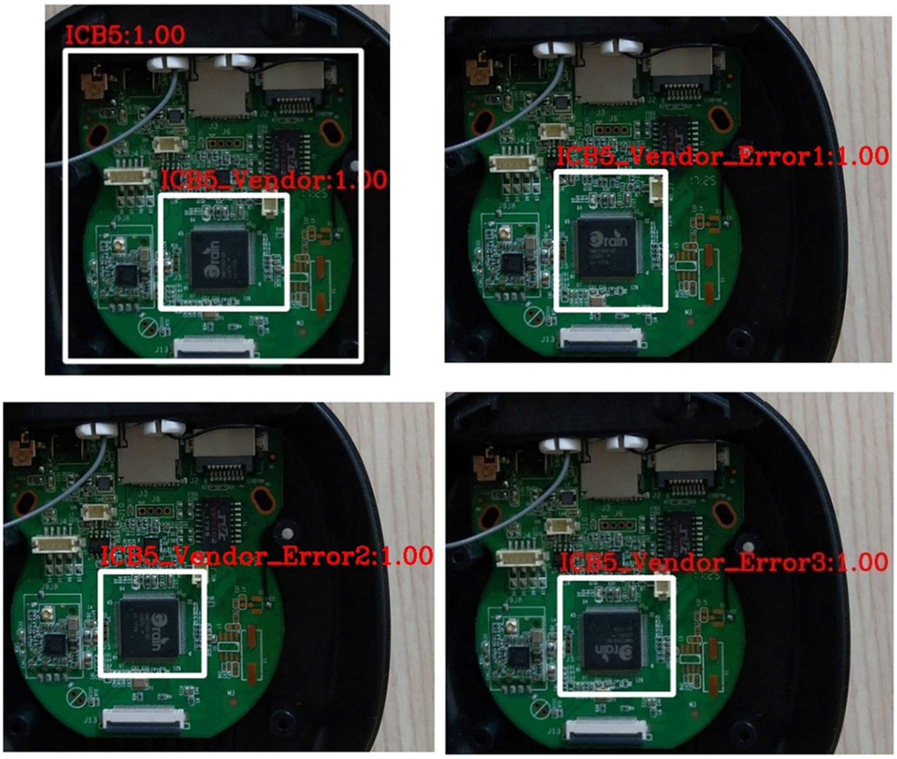


FIGURE 10 | Directionality classification and inspection of integrated circuit image (correct, error type 1, error type 2, and error type 3).

TABLE 4 | Confusion matrix of ICB Image directionality recognition results.

Original predict	Correct	Error 1	Error 2	Error 3	None
Correct	30	0	0	0	0
Error 1	0	6	0	0	0
Error 2	0	0	7	0	0
Error 3	0	0	0	6	0
None	0	0	0	1	0

manufacturing. In the third stage, the influence of the image augmentation fusion method on the identification results of the model is discussed based on the comparison results of the second stage. Finally, the fourth stage discusses the application of its model in the actual field. The results of the experiment show the following conclusions:

YOLOv3 Model Selection

In the experimental part of this study, because we hoped to apply the model to smart manufacturing and because the advantage of YOLO is the speed of image recognition, so we hoped to choose a model with excellent training cost and actual recognition results. After comparing YOLOv3, YOLOv3-tiny, YOLOv3-voc, YOLOv3-spp under the third version of YOLO, the experimental results show that YOLOv3-voc is the best choice, which can reach the highest 96.53% accuracy rate and 94.72% average accuracy rate during test stage under the experimental conditions, the performance is quite good. Although the second-place YOLOv3-tiny model also has an average accuracy of 93.56, the difference in training time to reach the same level is quite large, so the final selection of the model is YOLOv3-voc. Of course, if we further optimize various parameters or lengthen the overall training time, it is possible to obtain higher accuracy.

Effectiveness of Data Augmentation Methods

In the second model comparison, this study applied a data augmentation method to the dataset to increase the data size and learn more features. Among them, data augmentation methods include angle flipping, focus scaling, and image cropping. As a result, the size of the dataset increased from 100 images to 400 images. Thus, the original average accuracy rose from 91.66 to 94.72%, which proved that the model has a higher grasp of the image characteristics of the ICB after using the image augmentation fusion method.

Application of Directional Inspection of the Integrated Circuit Board

This study focuses on the actual image recognition of the ICB. We used the brand image of the ICB as the inspection target to determine the correctness of its installation direction. After experimental testing, a total of 160 images were used to complete the training test. In the last 50 test images, the detection accuracy rate reached 98%, exceeding the 90% threshold in general actual application environment, proving that the model could be used for application testing.

Discussion on the Number of Iterations of the YOLOv3_Tiny Model

This study also had a separate discussion on the YOLOv3_tiny model. The training cost of the model and the experimental data of the YOLOv3_tiny model are discussed in the first few subsections. Compared with other models, the training time is shorter due to its lightweight architecture. Although a high level of average test accuracy can be achieved through multiple training iterations, the overall time cost is still slightly higher than YOLOv3_voc. Nevertheless, its advantage is that the equipment is relatively standard, and it is easy to train a good model for application quickly.

CONCLUSIONS

Smart manufacturing must cover functions such as automated information perception, automated decision-making, and automated execution. What drives these automated processes rely on data and every piece of this data comes from various sensors, and image recognition is one of the methods that can be used. Moreover, based on the deep learning architecture, the work can be completed by the trained model. The results prove that YOLO's model can achieve the lowest model training cost in an automated environment that requires image recognition speed and excellent image recognition results using the ICB image under the pre-processing method of this study. Thus, the model is quite suitable for application in the smart manufacturing field, effectively achieving automatic perception.

This study also discusses several YOLO models. Among them, YOLOv3_voc has the best performance, with the highest accuracy rate of 91.9%. When combined with the pre-processing in experiment 2 of this study using the image data augmentation fusion method, the highest accuracy can reach 96.53, 4.5% higher than the original model without the data augmentation method. In the final experiment, the image of the ICB was used and the directional inspection accuracy could reach 98%, which met the 90% threshold required in general application. In addition, given the real-time nature of the production site, this study takes <1 s to identify each image, which can be a good candidate for application with real-time requirements. This proves the feasibility and accuracy of R-CNN in the field of smart manufacturing.

Regarding the research limitations in this study, since it is impossible to collect all different ICB image data, the ICB image data sources in this study are only specific to five different types of webcams. In addition, in terms of model selection, the YOLOv3 model was used in this study in consideration of both machine performance and accuracy. In the future, more innovative models and more various ICB image data can be used in this architecture. In addition, to optimize the parameters of this model for the future development of this study, the biggest problem is actually the availability of data. Although the R-CNN can achieve excellent image recognition results, it requires many data behind it and must be labeled as learning features. To achieve the ultimate automatic perception, automatic correction is needed. The automatic correction introduced in image recognition provides new data that can be imported into the dataset of the model for learning. If it could be improved, the results of the learning are believed to be more prominent. Another part is about the method of image pre-processing. Although this study uses image data augmentation fusion methods, it may be possible to import binary image processing to increase data in the future.

Finally, we hope the model can be applied to smart manufacturing as practical application to make overall learning adjustments. There will be some problems in the actual field, such as the effect of light that may cause reflections when the ICB image is automatically detected, resulting in unrecognizable results. Therefore, it may be necessary to sample the characteristics of the ICB itself and some other features to assist the image recognition process.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

S-YL contributed to conception, design, formal analysis, formulated methodology, funding acquisition of the study, and reviewed and edited the manuscript. S-YL and H-YL organized the data curation and wrote the first draft of the manuscript. H-YL analyzed the image data. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Ministry of Science and Technology of Taiwan under Grant MOST 109-2410-H-197-002-MY3.

ACKNOWLEDGMENTS

Authors thank Chi-Min Chang for assisting some part of the YOLO model implementation work and some part of the paper format editing.

REFERENCES

Bharati, P., and Pramanik, A. (2020). "Deep learning techniques—R-CNN to mask R-CNN: a survey," in *Computational Intelligence in Pattern Recognition* Advances in Intelligent Systems and Computing., eds A. K. Das, J. Nayak, B. Naik, S. K. Pati, and D. Pelusi (Singapore: Springer), 657–668. doi: 10.1007/978-981-13-9042-5_56

Bihi, T., Luwes, N., and Kusakana, K. (2018). Innovative quality management system for flexible manufacturing systems. in "*2018 Open Innovations Conference (OI)*" (Johannesburg). doi: 10.1109/OI.2018.8535610

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: optimal speed and accuracy of object detection. *arXiv [Preprint] arXiv:2004.10934.*. Available online at: http://arxiv.org/abs/2004.10934 (accessed August 20, 2021).

Diering, M., and Kacprzak, J. (2021). "Optical inspection software for a selected product on the smart factory production line," in *Advanced Manufacturing Processes II Lecture Notes in Mechanical Engineering*, eds V. Tonkonogyi, V. Ivanov, J. Trojanowska, G. Oborskyi, A. Grabchenko, I. Pavlenko, et al. (Cham: Springer International Publishing), 785–796. doi: 10.1007/978-3-030-68014-5_76

Gavrilescu, R., Zet, C., Foşalău, C., Skoczylas, M., and Cotovanu, D. (2018). "Faster R-CNN: an approach to real-time object detection," in *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)* (Iaşi), 0165–0168. doi: 10.1109/ICEPE.2018.8559776

Girshick, R. (2015). Fast R-CNN. *arXiv [Preprint] arXiv:1504.08083*. Available online at: http://arxiv.org/abs/1504.08083 (accessed August 18, 2021).

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv [Preprint] arXiv:1311.2524*. Available online at: http://arxiv.org/abs/1311.2524 (accessed August 19, 2021). doi: 10.1109/CVPR.2014.81

He, K., Zhang, X., Ren, S., and Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv [Preprint] arXiv:1406.4729*. doi: 10.1007/978-3-319-10578-9_23

Huang, Z., and Wang, J. (2019). DC-SPP-YOLO: dense connection and spatial pyramid pooling based YOLO for object detection. *arXiv [Preprint] arXiv:1903.08589*. Available online at: http://arxiv.org/abs/1903.08589 (accessed August 19, 2021).

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* 53, 5455–5516. doi: 10.1007/s10462-020-09825-6

Kimemia, J. G., and Gershwin, S. B. (1983). "An algorithm for the computer control of production in a flexible manufacturing system," in *1981 20th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes* (San Diego, CA), 628–633. doi: 10.1109/CDC.1981.269285

Kovrigin, E., and Vasiliev, V. (2020). Trends in the development of a digital quality management system in the aerospace industry. *IOP Conf. Ser. Mater. Sci. Eng.* 868:012011. doi: 10.1088/1757-899X/868/1/012011

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Red Hook, NY). Available online at: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed August 18, 2021).

Lian, J., Wang, L., Liu, T., Ding, X., and Yu, Z. (2021). Automatic visual inspection for printed circuit board via novel Mask R-CNN in smart city applications. *Sustain. Energy Technol. Assess.* 44:101032. doi: 10.1016/j.seta.2021.101032

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. *arXiv [Preprint] arXiv:1803.01534*. Available online at: http://arxiv.org/abs/1803.01534 (accessed August 20, 2021).

Maity, M., Banerjee, S., and Sinha Chaudhuri, S. (2021). "Faster R-CNN and YOLO based Vehicle detection: a survey" in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (Erode)* 1442–1447. doi: 10.1109/ICCMC51019.2021.9418274

Mukhopadhyay, A., Murthy, L. R. D., Arora, M., Chakrabarti, A., Mukherjee, I., and Biswas, P. (2019). "PCB Inspection in the Context of Smart Manufacturing," in *Research into Design for a Connected World Smart Innovation, Systems and Technologies*, ed A. Chakrabarti (Singapore: Springer), 655–663. doi: 10.1007/978-981-13-5974-3_57

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: unified, real-time object detection. *arXiv [Preprint] arXiv:1506.02640*. Available at: http://arxiv.org/abs/1506.02640 (accessed August 19, 2021).

Redmon, J., and Farhadi, A. (2016). YOLO9000: better, faster, stronger. *arXiv [Preprint] arXiv:1612.08242*. Available online at: http://arxiv.org/abs/1612.08242 (accessed August 19, 2021).

Redmon, J., and Farhadi, A. (2018). YOLOv3: an incremental improvement. *arXiv [Preprint] arXiv:1804.02767*. Available online at: http://arxiv.org/abs/1804.02767 (accessed August 19, 2021).

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: towards real-time object detection with region proposal networks. *arXiv [Preprint] arXiv:1506.01497*. Available online at: http://arxiv.org/abs/1506.01497 (accessed August 19, 2021).

Wang, C.-Y., Liao, H.-Y. M., Yeh, I.-H., Wu, Y.-H., Chen, P.-Y., and Hsieh, J.-W. (2019). CSPNet: a new backbone that can enhance learning capability of CNN. *arXiv [Preprint] arXiv:1911.11929*. Available online at: http://arxiv.org/abs/1911.11929 (accessed August 20, 2021).

Wang, T., Yao, Y., Chen, Y., Zhang, M., Tao, F., and Snoussi, H. (2018). Auto-sorting system toward smart factory based on deep learning for image segmentation. *IEEE Sens. J.* 18, 8493–8501. doi: 10.1109/JSEN.2018.2866943

Wang, Y., Liu, M., Zheng, P., Yang, H., and Zou, J. (2020). A smart surface inspection system using faster R-CNN in cloud-edge computing environment. *Adv. Eng. Inform.* 43:101037. doi: 10.1016/j.aei.2020.101037

Xu, B., Wang, W., Falzon, G., Kwan, P., Guo, L., Chen, G., et al. (2020). Automated cattle counting using Mask R-CNN in quadcopter vision system. *Comput. Electron. Agric.* 171:105300. doi: 10.1016/j.compag.2020.105300

Yadav, A., and Jayswal, S. C. (2018). Modelling of flexible manufacturing system: a review. *Int. J. Prod. Res.* 56, 2464–2487. doi: 10.1080/00207543.2017.1387302

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 15 November 2021
doi: 10.3389/fnbot.2021.762252

Multi-Modal Image Fusion Based on Matrix Product State of Tensor

Yixiang Lu, Rui Wang, Qingwei Gao, Dong Sun and De Zhu*

Anhui University, Hefei, China

Multi-modal image fusion integrates different images of the same scene collected by different sensors into one image, making the fused image recognizable by the computer and perceived by human vision easily. The traditional tensor decomposition is an approximate decomposition method and has been applied to image fusion. In this way, the image details may be lost in the process of fusion image reconstruction. To preserve the fine information of the images, an image fusion method based on tensor matrix product decomposition is proposed to fuse multi-modal images in this article. First, each source image is initialized into a separate third-order tensor. Then, the tensor is decomposed into a matrix product form by using singular value decomposition (SVD), and the Sigmoid function is used to fuse the features extracted in the decomposition process. Finally, the fused image is reconstructed by multiplying all the fused tensor components. Since the algorithm is based on a series of singular value decomposition, a stable closed solution can be obtained and the calculation is also simple. The experimental results show that the fusion image quality obtained by this algorithm is superior to other algorithms in both objective evaluation metrics and subjective evaluation.

Keywords: multi-modal, image fusion, tensor, matrix product state, singular value decomposition

OPEN ACCESS

Edited by:
Shin-Jye Lee,
National Chiao Tung University, Taiwan

Reviewed by:
Puming Wang,
Yunnan University, China
Jinjin Hai,
National Digital Switching System Engineering and Technological Research Centre, China

****Correspondence:***
Yixiang Lu
lyxahu@ahu.edu.cn

Received: *21 August 2021*
Accepted: *06 October 2021*
Published: *15 November 2021*

Citation:
Lu Y, Wang R, Gao Q, Sun D and Zhu D (2021) Multi-Modal Image Fusion Based on Matrix Product State of Tensor.
Front. Neurorobot. 15:762252.
doi: 10.3389/fnbot.2021.762252

1. INTRODUCTION

The purpose of image fusion is to synthesize multiple images of the same scene into a fusion image containing part or all information of each source image (Zhang, 2004). The fused image contains more information than each source image, thus, it is more suitable for machine processing and human visual perception. Image fusion has a wide range of applications in many fields, such as computer vision, remote sensing, medical imaging, and video surveillance (Goshtasby and Nikolov, 2007). The same type of sensors acquire information in a similar way, so the single-modal image fusion cannot provide information of the same scene from different aspects. On the contrary, multi-modal image fusion (Ma et al., 2019) realizes the complementarity of different features of the same scene through fusing the images collected by different types of sensors and generates an informative image for subsequent processing. As typical multi-modal images, infrared and visible images, CT and MRI images can provide distinctive features and complementary information, that is, infrared images can capture thermal radiation signal and visible images can capture reflected light signal; CT is mainly used for signal acquisition of sclerous tissue (e.g., bones), and MRI is mainly used for signal acquisition of soft tissue. Therefore, multi-modal image fusion has a wide range of applications in engineering practice.

To realize image fusion, many scholars have proposed a large number of fusion algorithms in recent years. In general, the fusion methods can be divided into two categories: the spatial-domain methods and the transform-domain methods. The typical methods in the first category include the weighted average method and principal component analysis (PCA) method (Yu et al., 2011) and so on. They fuse the gray values of image pixels directly. Although the direct operation on the pixels has low complexity, the fusion process is less robust to noise, and the results cannot meet the needs of the application in most cases. To overcome this shortcoming, a fusion method based on transform is proposed (Burt and Adelson, 1983; Haribabu and Bindu, 2017; Li et al., 2019). In general, the transform-based methods obtain the transformed coefficients of an image using a certain set of base functions, then fuse these coefficients through

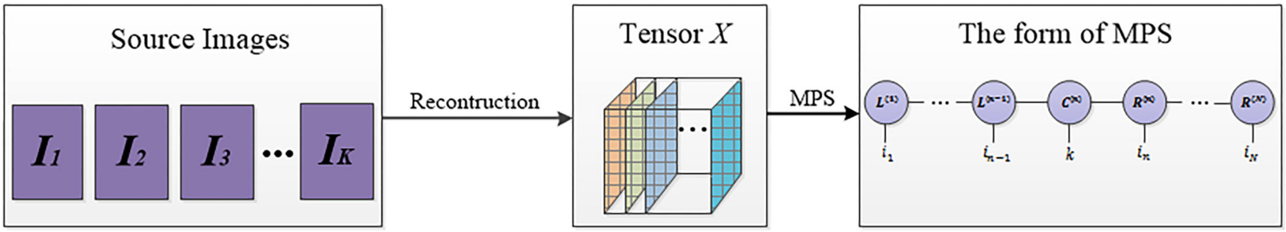


FIGURE 1 | The matrix product state (MPS) form of ***X***.

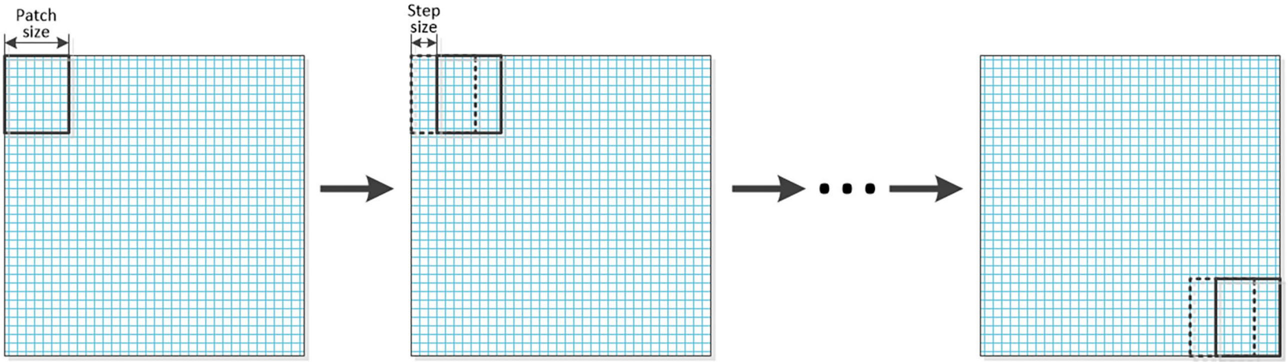


FIGURE 2 | The graphical representation of the sliding window technology.

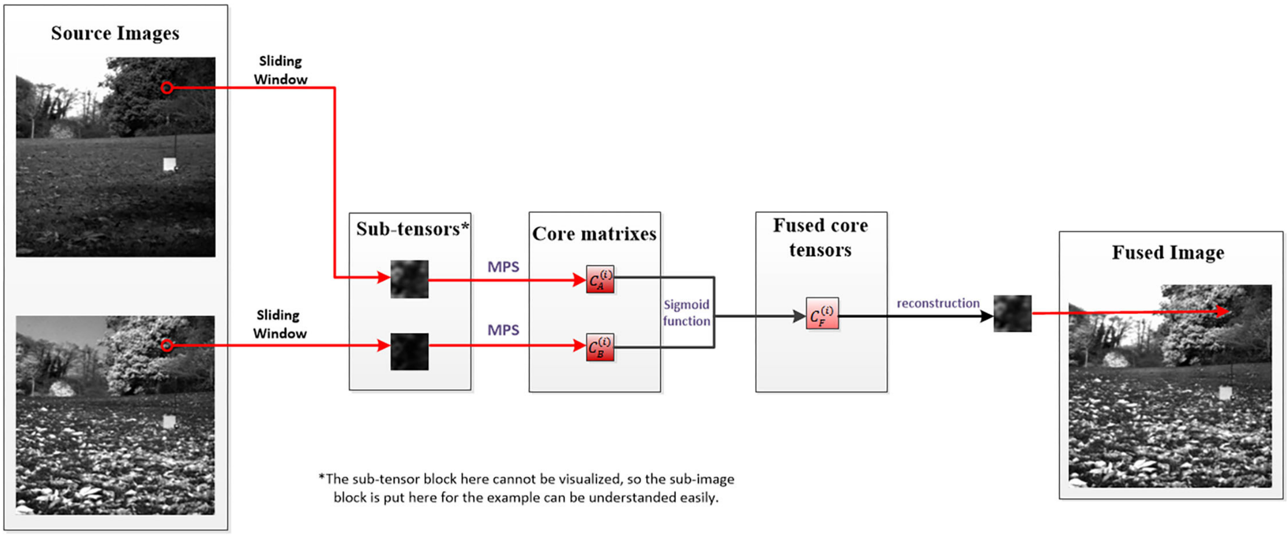


FIGURE 3 | An example for visualizing the process of decomposition and fusion.

certain fusion rules, and finally obtain the final fused image through the corresponding inverse transform. For example, Burt and Adelson (1983) formed a laplacian pyramid (LP) by desampling and filtering source images, and then designed different fusion strategies at each layer. Finally, the fused image is obtained by applying the inverse transform on the fusion coefficients. Haribabu and Bindu (2017) first decomposed the source images by using discrete wavelet transform (DWT) and fused the coefficients with predefined fusion rules, and then obtained the final image by applying the inverse discrete wavelet transform on fused coefficients. Because the transform-based method employs the average weighted fusion rules for the low-frequency components which carry the most energy of the image, there will be something wrong with the contrast loss of the final fused image.

In addition to traditional spatial-domain and transform-domain methods, sparse representation (SR) has been extensively used in image fusion in recent years (Yang and Li, 2010; Jiang and Wang, 2014; Liu et al., 2016; Zhang and Levine, 2016). The SR method assumes that the signal to be processed satisfies $y \in R^n$, then $y = Dx$, where $D \in R^{n\times m}(n << m)$ is an overcomplete dictionary, and n is the dimensions of the signal and m is the number of atoms in the dictionary D which is formed by a set of image subblocks, x is the sparse coefficients vector. The fused image is reconstructed by means of fusing the sparse coefficients. Although the SR-based method has achieved many results in the field of image fusion, some detailed information will be lost in the reconstructed image (e.g., the edges and textures tend to be smoothed), which limits the ability of the SR to express images (Yang and Li, 2010). To solve this problem, some scholars proposed some improved algorithms (Jiang and Wang, 2014; Liu et al., 2016). For instance, Jiang and Wang (2014) used morphological component analysis (MCA) to represent the source images more effectively. The MCA method first applied SR to separate the source images into two parts: cartoon and texture, then different fusion rules were designed to fuse these two parts respectively. Finally, a fused image with rich information was obtained, and more characteristic features of the source images were preserved.

As an extension of the vector and matrix, the tensor (Kolda and Bader, 2009) plays an important role in the high-dimensional data processing. In the field of computer science and technology, a tensor is a multi-dimensional array. It can be extended to some common data types, for example, a zero-order tensor can be defined as a constant, the tensor of order 1 is defined as a vector, the tensor of order 2 is defined as a matrix, the tensor of order 3 and the tensor of order N $(N \geq 3)$ is called high-order tensor. In essence, tensor decomposition is a high-order generalization of matrix decomposition, which is mainly applied to dimensionality reduction, sparse data filling, and implicit relationship mining. The information processing method based on tensor is more suitable for the processing of high-dimensional data and the extraction of feature information than vector and matrix, therefore, some relevant applications have been emerged in recent years (Bengua et al., 2015, 2017a,b; Zhang et al., 2016). In view of the excellent performance of tensors in representing

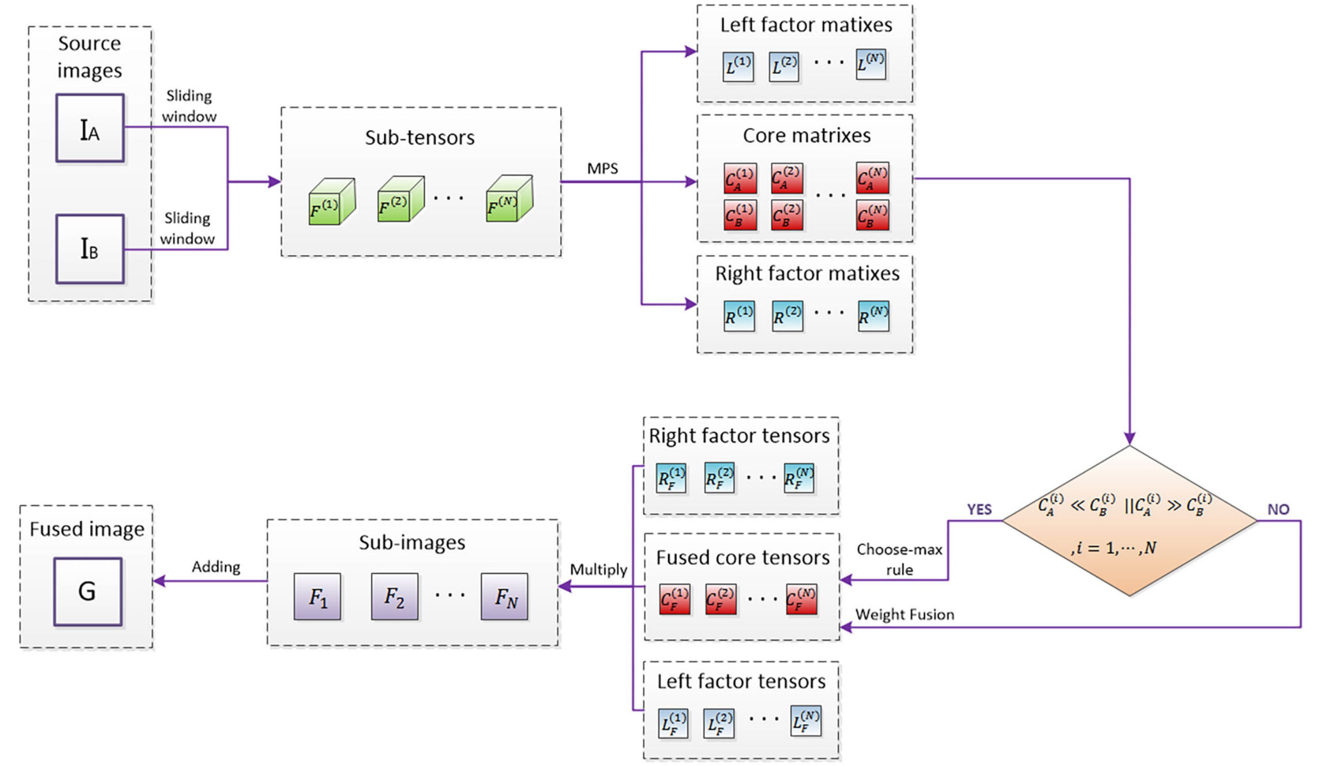

Source images
I_A
I_B
Sliding window
Sliding window
Sub-tensors
$F^{(1)}$ $F^{(2)}$ · · · $F^{(N)}$
MPS
Left factor matixes
$L^{(1)}$ $L^{(2)}$ · · · $L^{(N)}$
Core matrixes
$C_A^{(1)}$ $C_A^{(2)}$ · · · $C_A^{(N)}$
$C_B^{(1)}$ $C_B^{(2)}$ $C_B^{(N)}$
Right factor matixes
$R^{(1)}$ $R^{(2)}$ · · · $R^{(N)}$
$C_A^{(i)} \ll C_B^{(i)}\ ||C_A^{(i)} \gg C_B^{(i)}$
$,i = 1,\cdots,N$
YES
NO
Choose-max rule
Weight Fusion
Right factor tensors
$R_F^{(1)}$ $R_F^{(2)}$ · · · $R_F^{(N)}$
Fused core tensors
$C_F^{(1)}$ $C_F^{(2)}$ · · · $C_F^{(N)}$
Left factor tensors
$L_F^{(1)}$ $L_F^{(2)}$ · · · $L_F^{(N)}$
Multiply
Sub-images
F_1 F_2 · · · F_N
Adding
Fused image
G

FIGURE 4 | Fusion flowchart based on MPS.

high-dimensional data and feature extraction, a tensor-based high-order singular value decomposition method (HOSVD) (Liang et al., 2012) was applied to image fusion and achieved good results. In this method, the source image is initialized into a tensor which is subsequently decomposed into several sub-tensors by using a sliding window technique. Then, the HOSVD is applied on each sub-tensors to extract the corresponding features which are fused by employing certain fusion rules.

Since HOSVD is an approximate decomposition method, it will lead to the loss of information in the process of image fusion. At the same time, the calculation process is large and a stable closed-form solution cannot be obtained. To avoid loss of detailed information, a novel method based on matrix product state (MPS) is proposed to fuse the multi-modal images. Compared with HOSVD, MPS achieves the improvement of HOSVD and achieves the purpose of acquisition image information accurately. Moreover, being different from SR who linearly represents images by using atoms in an overcomplete dictionary, MPS decomposes image tensor into MPS. The main difference is that SR is approximate decomposition, while MPS is accurate decomposition. Therefore, in terms of signal reconstruction, MPS has better performance in signal expression. The main contributions of the article are outlined as follows: (i) Considering that image fusion depends more on local

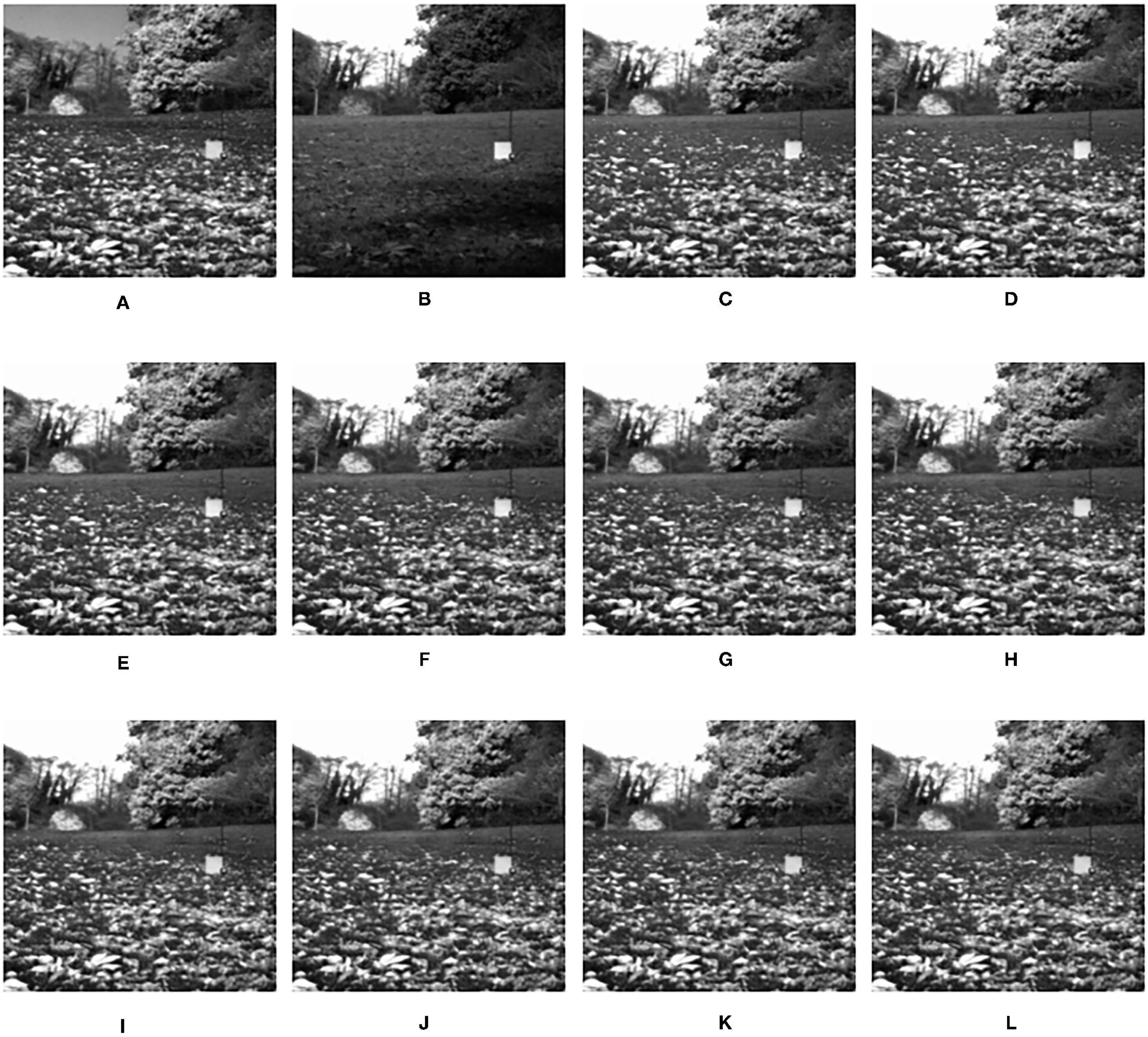


FIGURE 5 | The output fused images in patch size experiment. **(A)** original image (infrared image); **(B)** original image (visible image); **(C)** patch of size 2 × 2; **(D)** patch of size 4 × 4; **(E)** patch of size 6 × 6; **(F)** patch of size 8 × 8; **(G)** patch of size 10 × 10; **(H)** patch of size 12 × 12; **(I)** patch of size 14 × 14; **(J)** patch of size 16 × 16; **(K)** patch of size 18 × 18; **(L)** patch of size 20 × 20.

information of the source images and dividing the image into blocks can get more details of each pixel, the two source images are first divided into several sub-image blocks, and then the corresponding sub-image blocks are initialized into sub-tensors; (ii) We perform the MPS on each sub-tensor separately to obtain the corresponding core matrixes. The core matrixes are fused using the fusion rule based on the sigmoid function which incorporates the conventional choose-max strategy and the weighted average strategy. This fusion strategy can preserve the features of the multi-modal source images and reduce the loss of contrast to the greatest extent; (iii) Due to the application of MPS, the computational complexity of image fusion based on tensor is reduced dramatically. Hence, MPS decomposition is realized by computing a series of sub-tensors with maximum order 3. Moreover, a stable closed-form solution can also be obtained in the proposed algorithm.

The rest of the article is organized as follows. Section 2 introduces the theory of matrix product decomposition. In section 3, the algorithm principle and the fusion steps are detailly discussed. Subsequently, the results of the experiments are presented in section 4. Finally, some conclusions are drawn in section 5.

TABLE 1 | The influence of patch size.

Patch size	*SD*	*MI*	*SSIM*	Q_G	Q_P
2 × 2	63.9385	**0.9289**	0.6683	0.6543	0.7358
4 × 4	64.3397	0.8837	0.6680	0.6589	0.7677
6 × 6	64.5772	0.8810	0.6679	0.6658	0.7819
8 × 8	64.7225	0.8850	0.6678	0.6710	0.7893
10 × 10	64.8229	0.8835	0.6681	0.6738	0.7903
12 × 12	64.9079	0.8867	0.6686	0.6760	0.7900
14 × 14	64.9586	0.8846	0.6695	0.6783	0.7907
16 × 16	65.0244	0.8811	**0.6699**	**0.6813**	**0.7915**
18 × 18	66.0479	0.8765	0.6543	0.6663	0.7574
20 × 20	**66.1362**	0.9043	0.6532	0.6679	0.7573

Bold values mean maximum value of the same metrices in the same group of comparative experiments.

2. MPS FOR TENSOR

2.1. Tensor

Tensor is a generalization of the vector. A vector is a kind of tensor with order 1. For simplicity and accuracy of the following

TABLE 2 | The influence of step size.

Step size	*SD*	*MI*	*SSIM*	Q_G	Q_P
1	65.0244	0.8811	**0.6699**	**0.6813**	**0.7915**
2	65.0206	0.8832	0.6697	0.6809	0.7910
4	65.0283	0.8905	0.6690	0.6775	0.7888
6	**65.0316**	0.9087	0.6673	0.6751	0.7835
8	65.0304	0.9223	0.6666	0.6753	0.7811
10	64.1665	**0.9461**	0.6630	0.6778	0.7752

Bold values mean maximum value of the same metrices in the same group of comparative experiments.

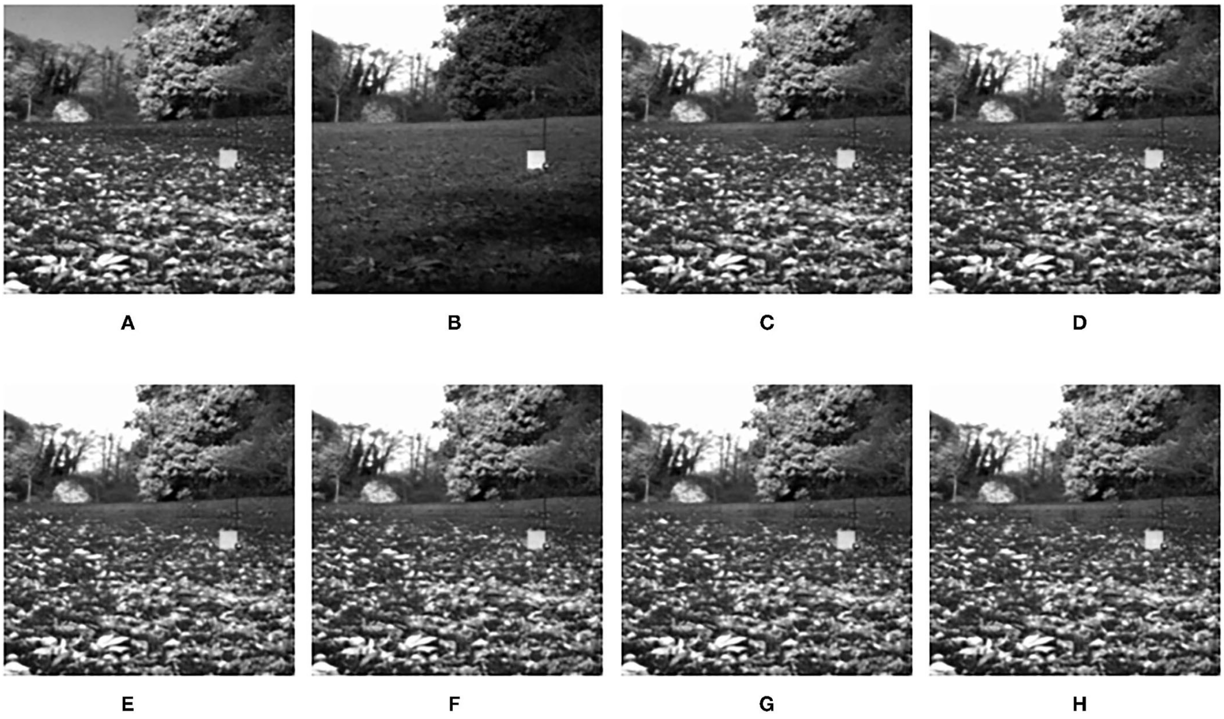


FIGURE 6 | The output fused images in step size experiment. **(A)** original image (infrared image); **(B)** original image (visible image); **(C)** step size = 1; **(D)** step size = 2; **(E)** step size = 4; **(F)** step size = 6; **(G)** step size = 8; **(H)** step size = 10.

expressions, first, we introduce some notations about tensors. The tensor of order 0 is a constant, represented by lowercase letter x; the tensor of order 1 is a vector represented by a bold lowercase letter $\mathbf{x}$; the tensor of order 2 is a matrix represented by a bold capital letter $\mathbf{X}$; the tensor of order 3 is a tensor represented by bold capital letters in italics $\boldsymbol{X}$. In this way, a tensor of order N and the size of each dimension are $I_1 \times I_2 \times \cdots \times I_N$ can be expressed as $\boldsymbol{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$, where I_i corresponds to the length of the i-th dimension. In general, we use $x_{i_1} \cdots x_{i_N}$ to represent the $(i_1, \cdots, i_N)$-th element of $\boldsymbol{X}$.

2.2. MPS for Tensor

The MPS decomposition (Perez-Garcia et al., 2006; Schollwock, 2011; Schuch et al., 2011; Sanz et al., 2016) aims to decompose an N-dimensional tensor $\boldsymbol{X}$ into the corresponding left-right orthogonal factor matrix and a core matrix. First, all the dimensions of an N-dimensional tensor $\boldsymbol{X}$ are rearranged, which lets the dimension K corresponding to the number of images to be fused, for example, if the number of source images is equal to 2, then $K = 2$. Additionally, the tensor $\boldsymbol{X}$ satisfies $\boldsymbol{X} \in R^{I_1 \times \cdots \times I_{n-1} \times K \times I_n \times \cdots \times I_N}$, $I_1 \geq \cdots \geq I_{n-1}, I_n \leq \cdots \leq I_N$, then the elements in the tensor $\boldsymbol{X}$ can be expressed in the form of MPS, and the schematic diagram of MPS form of $\boldsymbol{X}$ is shown in **Figure 1**:

$$x_{i_1 \cdots k \cdots i_N} = x^{(k)}_{i_1 \cdots i_n \cdots i_N} \approx \mathbf{L}^{(1)}_{i_1} \cdots \mathbf{L}^{(n-1)}_{i_{n-1}} \mathbf{C}^{(n)}_{k} \mathbf{R}^{(n+1)}_{i_n} \cdots \mathbf{R}^{(N+1)}_{i_N}. \tag{1}$$

$\mathbf{L}^{(j)}_{i_j}$ and $\mathbf{R}^{(j)}_{i_{(j-1)}}$ mentioned in the above formula are called left-right orthogonal factor matrix with size $\delta_{j-1} \times \delta_j$, where $\delta_0 = \delta_{N+1} = 1$, and they are all orthogonal:

$$\sum_{i_j=1}^{I_j} (L^{(j)}_{i_j})^T \mathbf{L}^{(j)}_{i_j} = \mathbf{I}, \quad (j = 1, \cdots, n-1) \tag{2}$$

and

$$\sum_{i_{j-1}=1}^{I_{j-1}} \mathbf{R}^{(j)}_{i_{j-1}} (R^{(j)}_{i_{j-1}})^T = \mathbf{I}, \quad (j = n+1, \cdots, N+1), \tag{3}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{C}^{n}_{k}$ is called core matrix.

TABLE 3 | Computation times of different algorithms.

Methods	Times (s)
DWT	0.1796
LP	0.3812
SR	6.4527
DTCWT − SR	3.8822
VGG	2.5067
MPS	1.8357

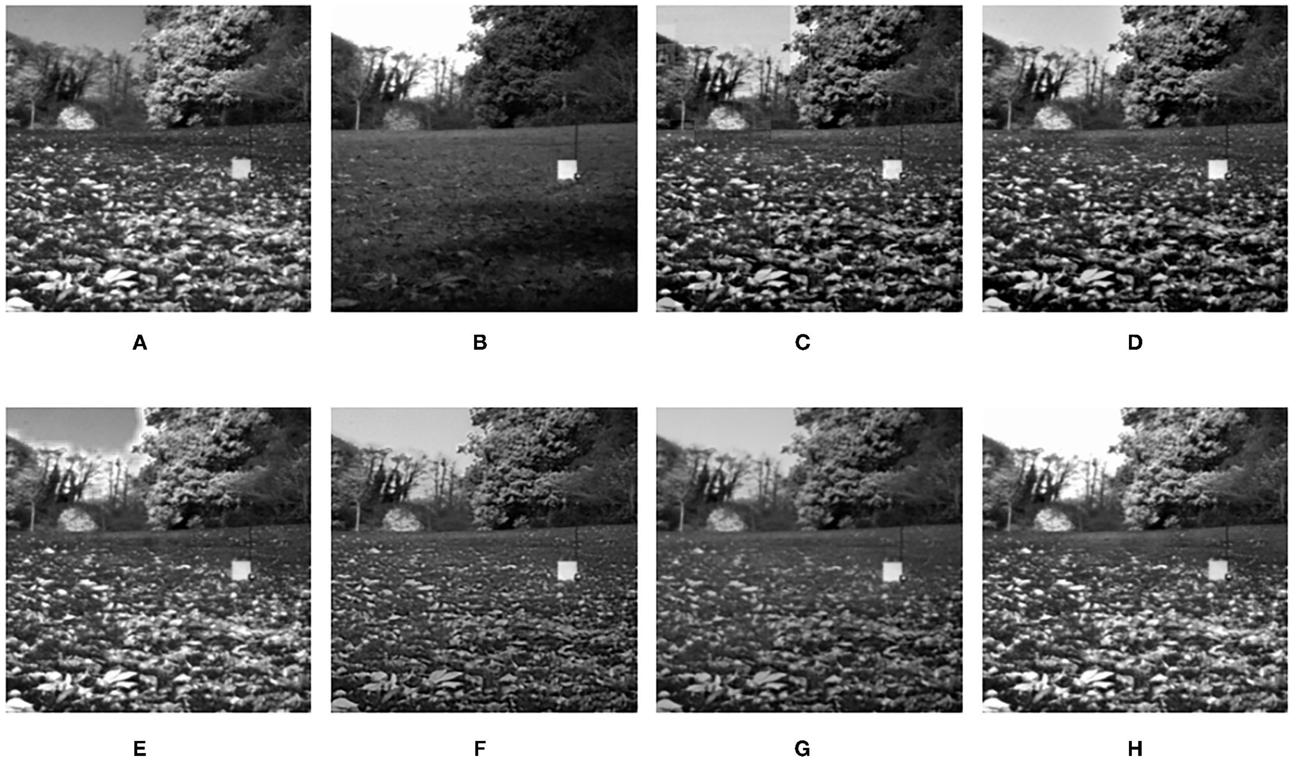


FIGURE 7 | Comparison experimental results of infrared and visible images. **(A)** original figure (infrared image); **(B)** original figure (visible image); **(C)** discrete wavelet transform (DWT); **(D)** laplacian pyramid (LP); **(E)** sparse representation (SR); **(F)** Dual-tree complex wavelet transform-sparse representation (DTCWT-SR); **(G)** VGG; **(H)** Matrix product state (MPS).

A tensor $\boldsymbol{X}$ can be decomposed into the form of (1) through two series of SVD decomposition. The process includes a left-to-right sweep and a right-to-left sweep. We summarize it in **Algorithm 1**.

```
Algorithm 1: Feature Extraction based on MPS
Input:
$\boldsymbol{X} \in R^{I_1\times\cdots\times I_{n-1}\times K\times\cdots\times I_N}$
Main Procedure:
Set $\mathbf{W}^{(1)} = \mathbf{X}_{(1)}$;
for $j = 0, 1, \ldots, n-1$ do
    $\mathbf{W}^{(j)} = \mathbf{USV}$;
    Reshape $\mathbf{U}^{(j)}$ to$\boldsymbol{U}$;
    $\mathbf{L}^{(j)}_{i_j} = \boldsymbol{U}(:, i_j, :)$;
end for
Reshape $\mathbf{V}^{(n-1)}$ to$\mathbf{W}^{N} \in R^{(\Delta_{n-1}K\cdots I_N)\times I_N}$;
for $j = N, N-1, \ldots, n$ do
    $\mathbf{W}^{(j)} = \mathbf{USV}$;
    Reshape $\mathbf{V}^{(j)}$ to $\boldsymbol{V}$;
    $\mathbf{R}^{(j+1)}_{i_{j-1}} = \boldsymbol{V}(:, i_{j-1}, :)$;
end for
Reshape $\mathbf{U}^{(n)}$into $\boldsymbol{C} \in R^{(I_{n-1}K\cdots I_N)\times I_N}$;
Set $\mathbf{C}^{n}_{k} = \boldsymbol{C}(:, k, :)$.
Output:
Core Matrix: $\mathbf{C}^{n}_{k} \in R^{\Delta_{n-1}\times\Delta_n}, k = 1, \cdots, K$;
Left Factor Matrix: $\mathbf{L}^{(j)}_{i_j}(i_j = 1, \cdots, I_j, j = 1, \cdots, n-1)$;
Right Factor Matrix: $\mathbf{R}^{(j)}_{i_{(j-1)}}(i_{(j-1)} = 1, \cdots, I_{(j-1)}, j = n+1, \cdots, N+1)$
```

3. IMAGE FUSION BASED ON MPS

In this section, the whole process of image fusion will be described. The source images which have been reconstructed into tensors are decomposed into a series of sub-tensors by using the sliding window technology. The graphical representation of the sliding window technology is shown in **Figure 2**. Then MPS is applied to the decomposed sub-tensors to obtain the core matrixes, and the sigmoid function is used for the fusion of each pair of core matrixes to obtain the fused core matrixes.

The specific theoretical concepts of decomposition and fusion are described in sections 3.1, 3.2, respectively, and the overall process of image fusion proposed in this article is described in section 3.3.

3.1. Tensor Decomposition by MPS

For the two source images A and B with sizes of $M \times N$, we use them to construct a tensor X with dimension $M \times N \times 2$. Taking into account the importance of local information of the

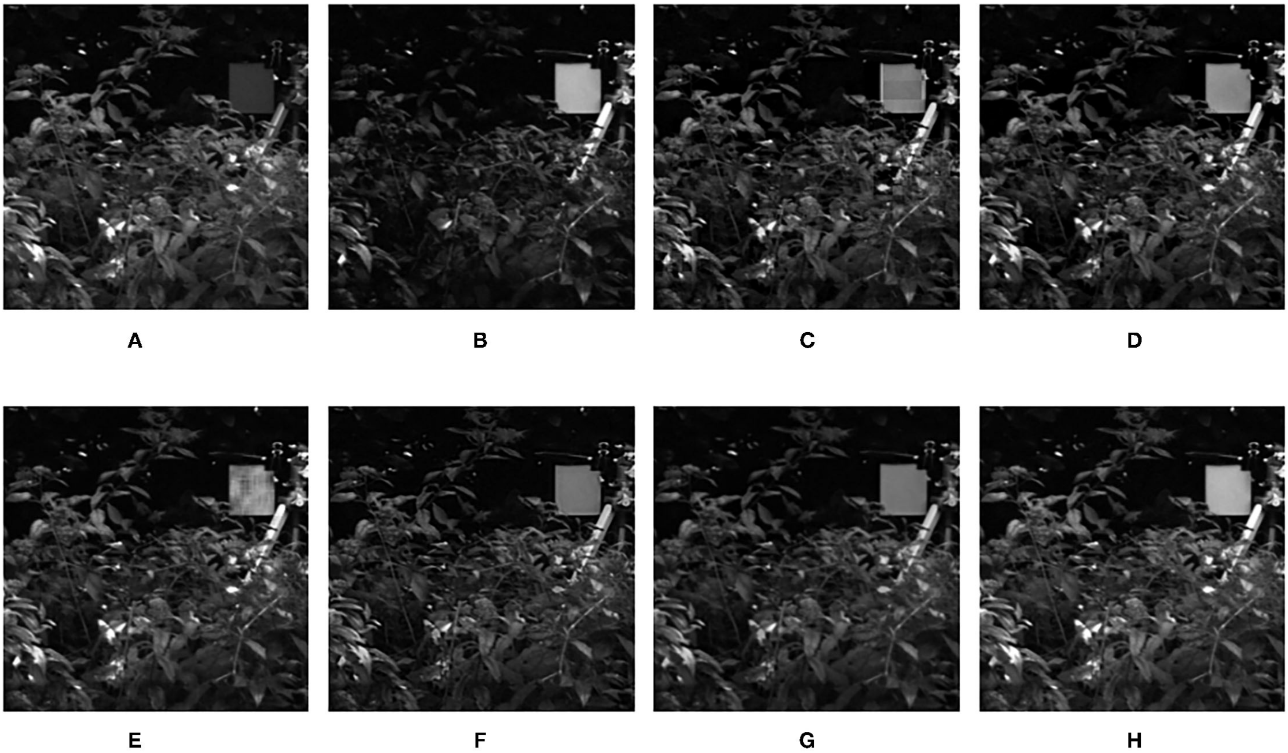


FIGURE 8 | Comparison experimental results of infrared and visible images. **(A)** original image (infrared image); **(B)** original image (visible image); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

source image, a sliding window technology is used to decompose it into several sub-tensors $\boldsymbol{F}$ with dimension $\overline{M} \times \overline{N} \times 2$, and the sliding step p used should satisfy $p \leq \min\{\overline{M}, \overline{N}\}$; the sub-tensor $\boldsymbol{F}$ is obtained by the **Algorithm 2**, as follows. In **Algorithm 2**, the fix$((M - patch\ size)/stepsize)$ represents the nearest integer to $(M - patch\ size)/stepsize$ and fix$((N - patch\ size)/stepsize)$ represents the nearest integer to $(N - patch\ size)/stepsize$. Then, MPS is applied to each of the sub-tensors.

Algorithm 2: The Sub-tensor obtained by Sliding Window Technology

Input:
$\boldsymbol{X} \in R^{M \times N \times 2}$

Main Procedure:

```
for $i = 1, 1 + stepsize, \ldots, 1 + stepsize \times fix(\frac{M - patch\ size}{stepsize})$ do
    for $j = 1, 1 + stepsize, \ldots, 1 + stepsize \times fix(\frac{N - patch\ size}{stepsize})$ do
        $\mathbf{F} = \boldsymbol{X}(i : i + patch\ size - 1, j + patch\ size - 1, :)$;
    end for
end for
```

Output:
sub-tensor: $\mathbf{F} \in R^{\overline{M} \times \overline{N} \times 2}$;

3.2. Design of Fusion Rule

We introduce the sigmoid function as the fusion rule of the characteristic coefficients, the fusion coefficient of each core matrix can be defined as follows:

$$e_i(l) = \sum_{m=1}^{\bar{M}} \sum_{n=1}^{\bar{N}} |\mathbf{C}_i(m, n, l)| \qquad l = 1, 2 \tag{4}$$

where the subscript i indicates the number of each sub-image, and l is the label of the corresponding source image.

For $e_i(l)$ obtained in the previous section, the fusion rule is selected by comparing the values of $e_i(1)$ and $e_i(2)$. When $e_i(1)$ is much less or much more than $e_i(2)$, we use the Max rule, and when the relationship between $e_i(1)$ and $e_i(2)$ satisfy the other relation, we use weighted fusion to fuse the corresponding coefficient matrix and then get the final fusion coefficient matrix. The function is as follows:

$$\begin{aligned} D_i = & \frac{1}{1 + exp(-kln(\frac{e_i(1)}{e_i(2)}))} \\ & \times \mathbf{C}_i(:, :, 1) + \frac{exp(-kln(\frac{e_i(1)}{e_i(2)}))}{1 + exp(-kln(\frac{e_i(1)}{e_i(2)}))} \times \mathbf{C}_i(:, :, 2) \end{aligned} \tag{5}$$

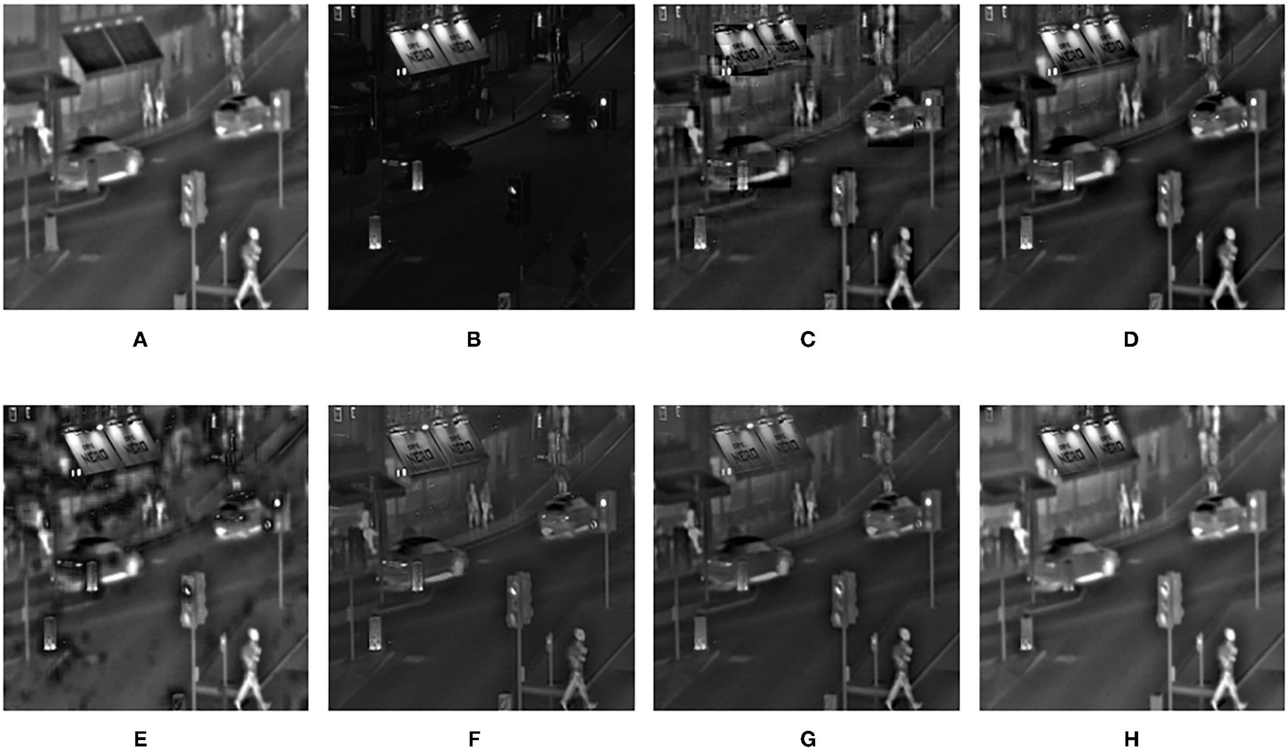


FIGURE 9 | Comparison experimental results of infrared and visible images. **(A)** original image (infrared image); **(B)** original image (visible image); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

where k is the shrinkage factor of the mentioned sigmoid function. After $\mathbf{D}_i$ is obtained, each of the fused sub-image blocks F_i can be reconstructed by the inverse operations of MPS. Then the sub-image blocks F_i is used to obtain the final fused image G.

To make the process of decomposition and fusion more concrete, the first group of the experiment images is used as an example to make a flowchart as shown in **Figure 3**:

3.3. The Process of Image Fusion Based on MPS

The process of image fusion based on MPS can be divided into the seven steps as follows

1. Input two source images;
2. Reconstructed the two source images into a third-order tensor, and the sub-tensors are extracted by sliding window technology;
3. Matrix product state decomposition is used on sub-tensors to obtain left and right factor matrixes and core matrixes;
4. Compare the vectors representing source image 1 and source image 2 in the core matrixes obtained in step 3, and obtain the fused matrixes by corresponding their quantitative relations to different situations of the sigmoid function, and then construct it as sub-tensors;
5. Multiply the fused sub-tensors by left and right factor tensors to obtain sub-images;
6. Sub-images addition;
7. Output fused image.

The specific flowchart is shown in **Figure 4**.

4. EXPERIMENTS

4.1. Objective Evaluation Metrics

1. Standard deviation (SD)
   SD is defined as follows:

$$SD = \sqrt{\frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} (F(x,y) - \mu)^2}, \tag{6}$$

   where μ is the average value of the fused image, H and W are the length and width of the image, respectively. SD is mainly used to measure the contrast of the fused image.
2. Mutual information (MI)
   Mutual information is defined as follows:

$$MI = \sum_{x=1}^{L} \sum_{y=1}^{L} h_{R,F}(i,j) log_2 \frac{h_{R,F}(i,j)}{h_R(i) + h_F(j)}, \tag{7}$$

   where $h_{R,F}(i,j)$ is the normalized joint distribution gray histogram between the source image R and the fused image

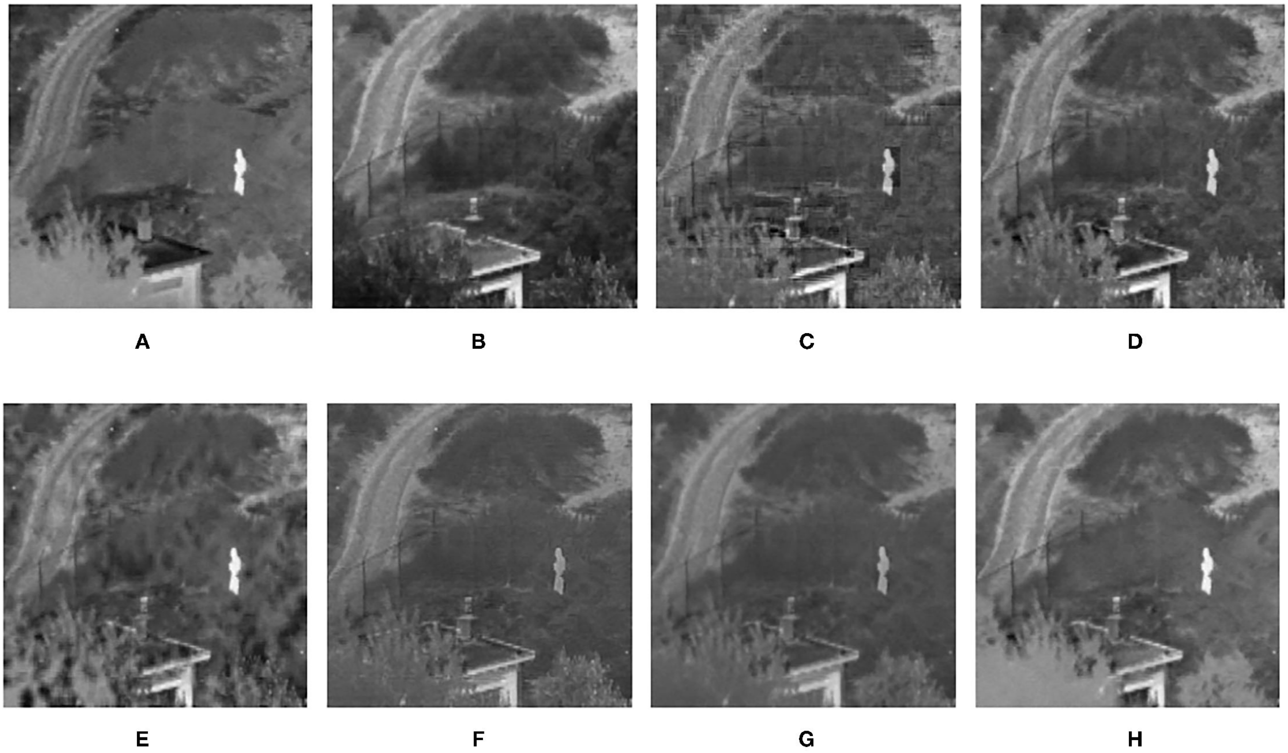


FIGURE 10 | Comparison experimental results of infrared and visible images. **(A)** original image (infrared image); **(B)** original image (visible image); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

F, $h_R(i)$ and $h_F(j)$ are the normalized marginal distribution histogram of the two source images, respectively, L is the number of gray levels.

3. Structural similarity (SSIM)
   Structural similarity is defined as follows:

$$SSIM(x,y) = (\frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + 1})^{\alpha}(\frac{2\sigma_x\sigma_y + c_1}{\sigma_x^2 + \sigma_y^2 + 1})^{\beta}(\frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3})^{\gamma}, \quad (8)$$

   where μ_x and μ_y are the average value of x and y. The middle term represents the similarity of contrast, σ_x and σ_y is the SD of x and y. The right term characterizes the structural similarity, and σ_{xy} is the covariance of x and y. The c_1, c_2, and c_3 are three constants, and the parameters α, β, and γ, respectively, adjust the contribution of the three terms. *SSIM* can calculate the similarity between the fused image and the source image. Its value which is between 0 and 1 is closer to 1, the more similar the two images are. The average value of the fused image and the two source images A and B is taken as the final evaluation metric, namely

$$SSIM = \frac{1}{2}(SSIM_A + SSIM_B). \quad (9)$$

4. Gradient based fusion metric (Q_G)
   Q_G is defined as follows:

$$Q_G = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W}(Q_{AF}(x,y)w_A(x,y) + Q_{BF}(x,y)w_B(x,y))}{\sum_{x=1}^{H}\sum_{y=1}^{W}(w_A(x,y) + w_B(x,y))}, \quad (10)$$

   where $Q_{AF}(x,y) = Q_{AF_g}(x,y)Q_{AF_\alpha}(x,y)$, at each pixel (x,y),$Q_{AF_g}(x,y)$ and $Q_{AF_\alpha}(x,y)$ denote the edge strength and orientation preservation values. $Q_{BF}(x,y)$ is defined as the same as $Q_{AF}(x,y)$. The weighting factors $w_A(x,y)$ and $w_B(x,y)$ indicate the significance of $Q_{AF}(x,y)$ and $Q_{BF}(x,y)$. Q_G is an important fusion image quality evaluation method computing the amount of gradient information that is injected into the fused image from the source image.

5. Phase congruency based fusion metric (Q_P)
   The Q_P is defined as follows:

$$Q_P = (P_p)^{\alpha}(P_M)^{\beta}(P_m)^{\gamma}, \quad (11)$$

   where p, M, and m refer to phase congruency, maximum, and minimum moments. The parameters α, β, and γ are set to 1 in this article. For more detailed information on parameters, please refer to the article Hong (2000). Q_P measures the extent that the salient features in the source image are preserved.

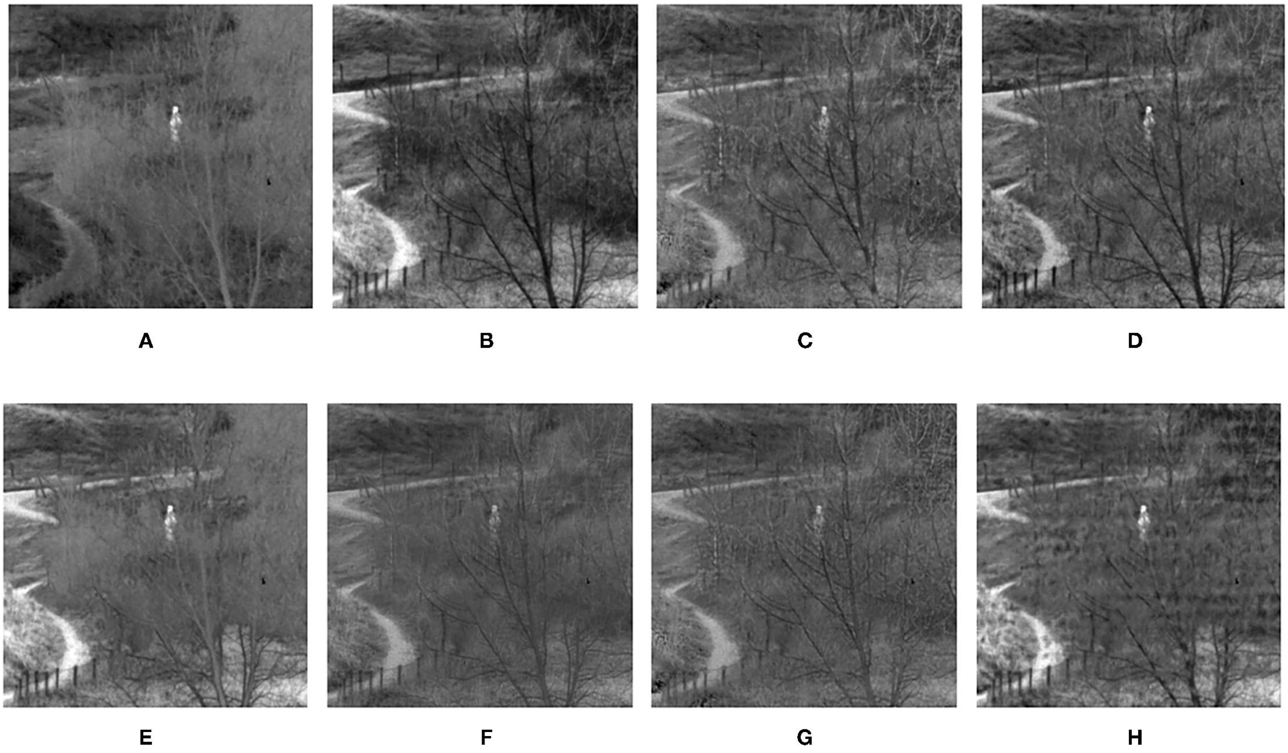


FIGURE 11 | Comparison experimental results of infrared and visible images. **(A)** original image (infrared image); **(B)** original image (visible image); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

4.2. Study of Patch Size and Step Size

Considering the sliding window technology is used, we will first study the respective influence of the size of the sub-image block and the step size of the sliding window on the performance of the fusion image experimentally. In the following statement we use patch size and step size to call the two factors briefly. To obtain the optimal patch size and step size, we will use a pair of infrared and visible images as source images, as shown in **Figures 5A,B**. In the experiment of patch size, the patch size is set to 2 × 2, 4 × 4, 6 × 6, 8 × 8, 10 × 10, 12 × 12, 14 × 14, 16 × 16, 18 × 18, and 20 × 20 with the step size fixed to 1 and shrinkage factor fixed to 200. In the experiment of step size, the step size is set to 1, 2, 4, 6, 8, and 10 with the patch size fixed to 16 × 16 and the shrinkage factor fixed to 200. The experimental results based on the objective evaluation metrics are shown in **Tables 1**, **2**. The output fused images are shown in **Figures 5**, **6**.

It can be seen clearly from **Table 1**, in most cases, that the best results can be obtained when the size of the sub-image block is 16 × 16. According to simple analysis, when the sub-image block is too small, the image characteristics cannot be effectively represented. Additionally, it can be seen from **Table 2** that when the step size is 1, the best result can be obtained. According to simple analysis, when the step size is too large, local information of the image may be lost or cannot be displayed well. Therefore, the in following experiments, the patch size was set to 16 × 16, and the step size was set to 1.

4.3. Computation Complexity

The computation time of each group of experimental images is recorded when different fusion algorithms are used. Experimental results show that the complexity of the proposed algorithm is lower than other algorithms. The results are shown in **Table 3** as follows:

All the codes are performed under MATLAB R2014a running on computer equipment with an Intel i7-7700K CPU (4.2 GHz) and 16 GB of RAM. As can be seen from the table, compared with SR and Dual-tree complex wavelet transform-sparse representation (DTCWT-SR), the running of the proposed algorithm is faster. In general, the computational complexity of the proposed algorithm is reduced.

4.4. Experimental Results and Discussion

In this section, the effectiveness of the proposed method is further verified by comparing the experimental results of this algorithm with other fusion methods. The comparison methods used are DWT (Haribabu and Bindu, 2017) and LP (Burt and Adelson, 1983), SR-based methods (Liu et al., 2016), VGG-Net (Hui et al., 2018), and DTCWT-SR (Singh et al., 2012). In addition to the infrared and visible images used in the previous section, CT and MRI medical images are also used for contrast experiments. The performance of each algorithm is evaluated by calculating the evaluation metrics based on the fusion results. In the experiment, all the experimental source image size is 256 × 256, the fixed patch

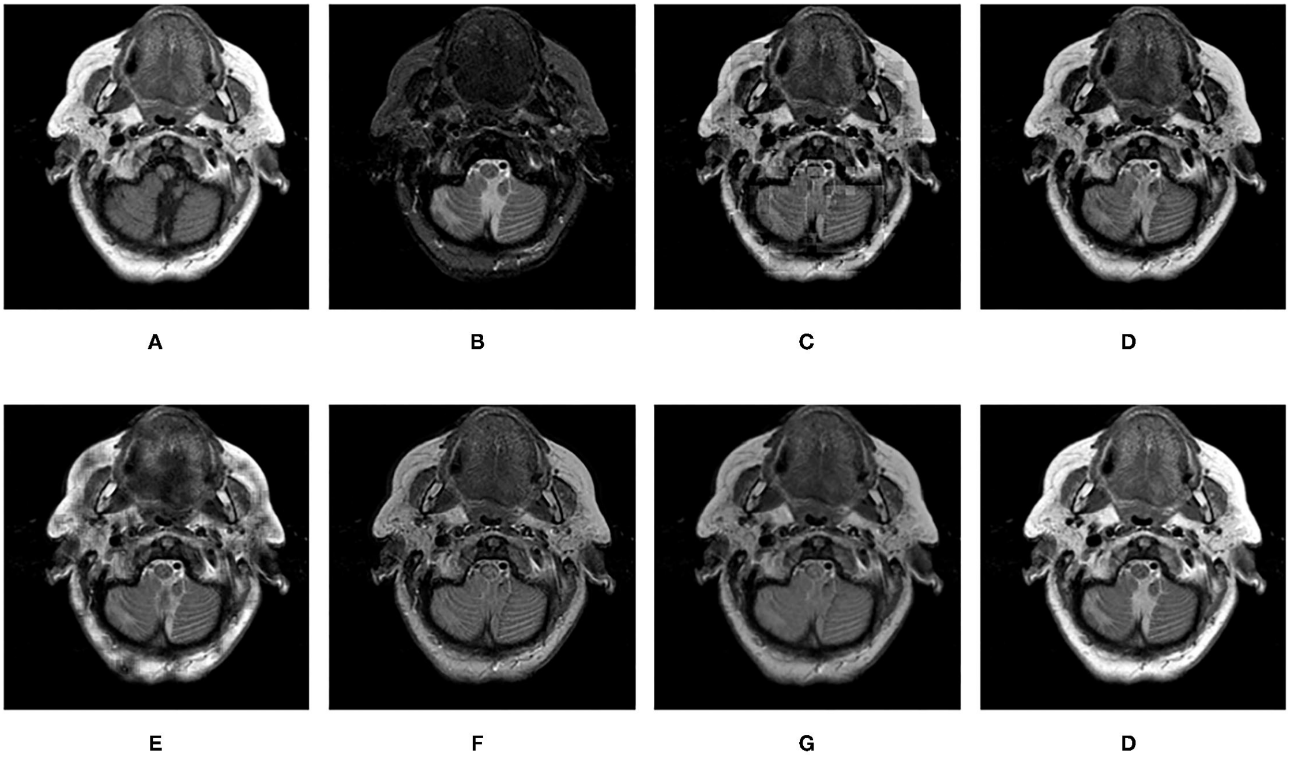


FIGURE 12 | Comparison experimental results of CT and MRI images. **(A)** original image (CT); **(B)** original image (MRI); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

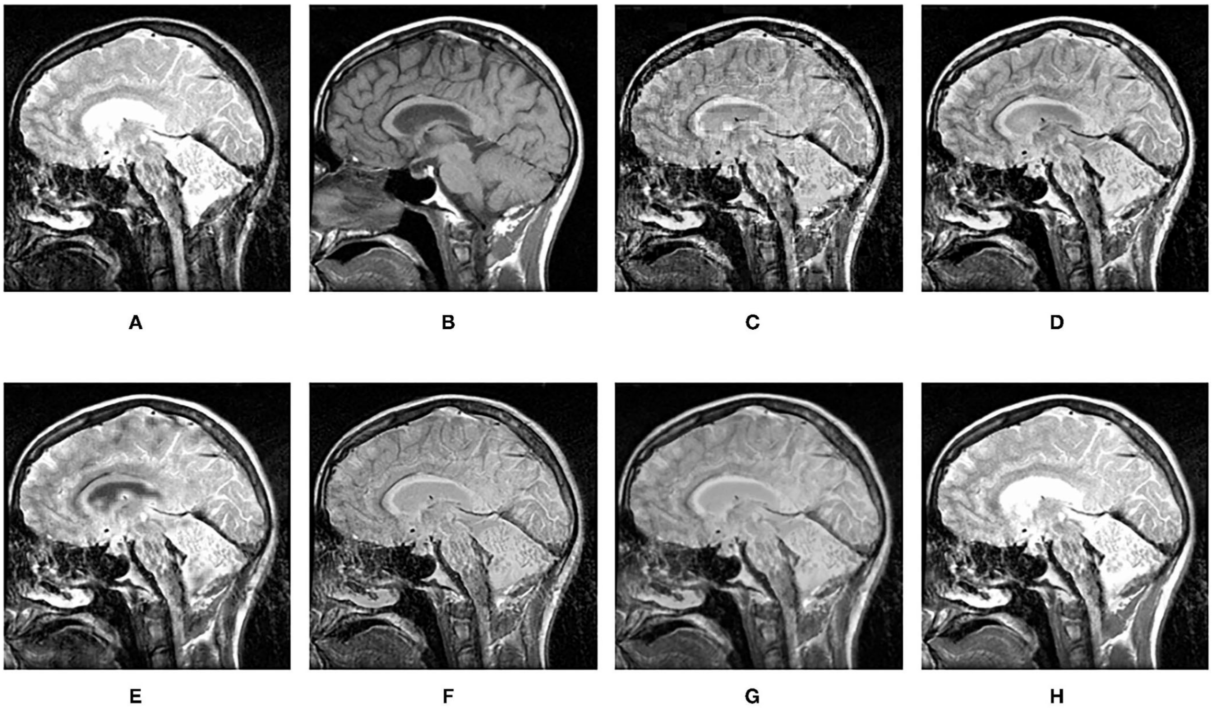


FIGURE 13 | Comparison experimental results of CT and MRI images. **(A)** original image (CT); **(B)** original image (MRI); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

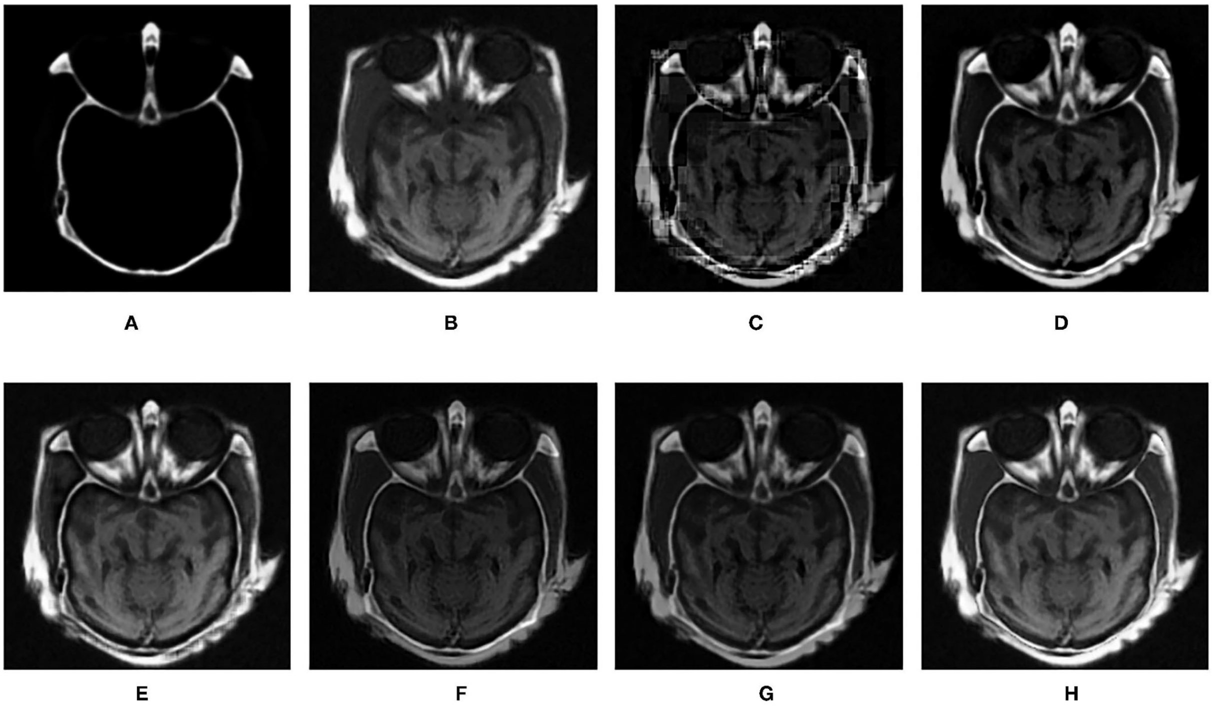


FIGURE 14 | Comparison experimental results of CT and MRI images. **(A)** original image (CT); **(B)** original image (MRI); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

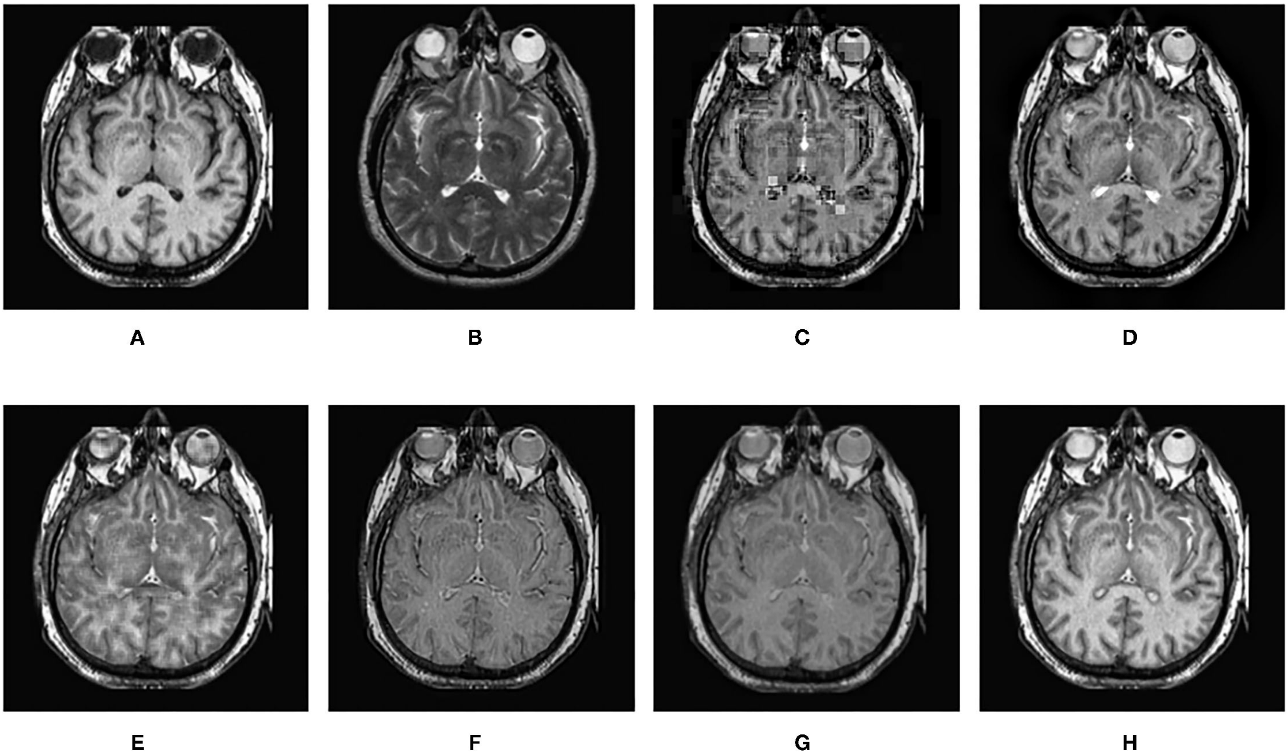


FIGURE 15 | Comparison experimental results of CT and MRI images. **(A)** original image (CT); **(B)** original image (MRI); **(C)** DWT; **(D)** LP; **(E)** SR; **(F)** DTCWT-SR; **(G)** VGG; **(H)** MPS.

size is 16 × 16, the step size is 1, and the shrinkage factor k is 200. The proposed method and several comparison algorithms are applied to nine pairs of source images. The experimental results are shown in **Figures 7–15**, respectively. The objective evaluation metrics values of the nine pairs of images are shown in **Tables 4**, **5**.

It can be seen from the table that in most cases, the algorithm proposed in this article can achieve optimal results, especially for CT and MRI images, the various metrics of the results of MPS are much higher than other methods. For infrared and visible images, the method in this article can also achieve optimal results under more than half of the evaluation metrics. These results show that the proposed method is better than other methods for multi-modal image fusion. This advantage mainly benefits from two aspects: (i) The sliding window method is adopted to divide the image into several sub-images, so the local information of the image can be captured well; (ii) MPS method is an accurate decomposition and reconstruction method, so in the process of image fusion, there will be no loss of information due to the solution.

Further analysis of the experimental results shows that: (i) On the whole, VGG-Net has the worst performance in all cases. Compared with other comparison methods, there is a big gap in various evaluation metrics. This is because the information captured is insufficient in the layer-by-layer feature extraction of the source image, and when the details of the fusion image are weighted by the final weight graph, the contrast of the initial detail part of the fusion image is reduced; (ii) Among the two multi-scale methods used, DWT fusion method performs poorly. This is because the DWT method is based on Haar wavelet to achieve fusion, which can only capture image features in horizontal and vertical directions but cannot capture more basic features of the image; LP method is better than the DWT method because the Laplacian pyramid generates only one band-pass component at each scale, which reduces the possibility of being affected by noise; (iii) The results obtained by SR method are better than other multi-scale methods in most cases but not as good as the proposed method. This is because the signal representation ability of SR is better than that of multi-scale transformation, and errors will occur in the process of signal reconstruction, which is unavoidable for the SR method. The method proposed in this article can effectively avoid this problem by non-destructive tensor reconstruction. In addition, the "max-L1" rule of direct fusion in the spatial domain will lead to spatial inconsistency, which affects the performance of the SR method; (iv) DTCWT-SR is an method that multi-scale method combined with SR method. By comparing the objective evaluation metrics, the fusion performance of the algorithm is better than SR in some aspects, but it is still poor compared with MPS.

In addition to objective evaluation, the performance of the algorithm in this article is also discussed through some visual comparisons of the fused images. In general, the

TABLE 4 | Comparison of objective evaluation metrics of infrared and visible images.

Figure	Metrics	DWT	LP	SR	DTCWT-SR	VGG	MPS
Figures 7A,B	*SD*	57.3495	59.1760	55.0023	65.0175	49.1788	**65.0244**
	MI	0.3915	0.3981	0.7197	0.5625	0.4373	**0.8811**
	SSIM	0.6308	0.6481	0.6602	0.6025	0.6266	**0.6699**
	Q_G	0.6573	0.6990	**0.7014**	0.6874	0.5975	0.6813
	Q_P	0.6769	0.7765	0.7392	0.7586	0.7139	**0.7916**
Figures 8A,B	*SD*	37.3080	39.6529	38.5454	41.7521	32.5971	**42.1726**
	MI	0.5262	0.5779	0.6457	0.5969	0.6614	**0.9720**
	SSIM	0.7674	0.7688	**0.8170**	0.8013	0.7603	0.8017
	Q_G	0.5558	0.6168	0.5753	0.6094	0.5995	**0.6748**
	Q_P	0.6769	0.7764	0.6895	0.7632	0.7659	**0.8349**
Figures 9A,B	*SD*	31.6360	35.0567	35.3218	**36.0795**	23.3351	35.0392
	MI	0.3357	0.4055	0.4509	0.3861	0.4190	**0.9292**
	SSIM	0.5701	0.6088	0.5350	0.4449	0.4284	**0.6157**
	Q_G	0.5755	0.6560	0.4487	0.5945	0.5352	**0.7249**
	Q_P	0.4050	0.5384	0.2579	0.5002	0.5816	**0.7004**
Figures 10A,B	*SD*	28.3593	29.2224	28.5693	29.7191	22.83747	**29.8865**
	MI	0.2238	0.2356	0.2720	0.2582	0.2604	**0.7066**
	SSIM	0.6288	**0.7088**	0.6331	0.4864	0.5494	0.6926
	Q_G	0.3999	0.4835	0.3304	0.4211	0.3482	**0.5204**
	Q_P	0.1892	0.2996	0.1220	0.2577	0.2541	**0.3927**
Figures 11A,B	*SD*	23.9236	25.6275	31.2000	**39.5077**	29.1652	31.7625
	MI	0.1528	0.1573	0.2883	0.4184	0.3954	**0.6704**
	SSIM	0.4223	0.4544	0.4624	0.4025	0.4351	**0.4911**
	Q_G	0.4217	0.5184	0.3987	0.5021	0.3749	**0.5204**
	Q_P	0.2256	0.3745	0.1827	0.3612	0.2683	**0.4304**

Bold values mean maximum value of the same metrices in the same group of comparative experiments.

proposed method achieves the best visual effect among all the fusion images.

The fusion results of infrared-visible images are shown in **Figures 7–11**. It can be seen from the figure that the method proposed in this article has good adaptability, and the fusion images are obtained to retain the information of the infrared and visible images, respectively. In **Figure 7**, both the multi-scale fusion method and SR show varying degrees of artificial traces at the junction between the trees and the sky in the upper left corner, while DTCWT-SR and VGG-Net resulted in severe contrast loss. In **Figure 8**, the white squares in infrared picture are dimming in varying degrees in DWT, LP, SR, DTCWT-SR, and VGG-Net methods, and the leaf luster in the visible image is not well-displayed in the VGG-Net method. In **Figure 9**, DWT and SR show the phenomenon of information loss. LP, DTCWT-SR, and VGG can get relatively complete fusion images, but the brightness is weaker than MPS. The clarity of the billboard in the upper left corner of the fused image is better in the MPS method. In **Figure 10**, the fused images obtained by DWT and SR show some small black blocks, that is information loss, while the human shape brightness on the right side of the images obtained by LP, DTCWT-SR, and VGG method is low. The reason for these shortcomings is the fusion rules used in the fusion process all have a certain degree of weighting on the source image. Our fusion rules based on the sigmoid function can well avoid these shortcomings, that is, in the image, whose colors are only black and white, the weight of the white part of the image will be much larger than that of the black part, thus, evolving into the Choose-max rule. In **Figure 11**, compared with the other five comparison methods, it can be seen that the human figure on the right and the branch on the lower right corner of the fusion image obtained by MPS have the highest resolution.

Figures 12–15 are the fusion results of CT and MRI medical images. It can be seen from the experimental results that the DWT method cannot to be applied to the fusion of medical images, and the other four methods can obtain a complete image. In **Figure 12**, LP, DTCWT-SR, and VGG-Net methods have no loss in details, but the sharpness of the light and dark junction is insufficient, the edge is blurred, and the contrast is lost. However, the bottom of the fused image obtained by the SR method is fractured, indicating that there is information loss. In **Figure 13**, the spine in the lower right corner and the jaw in the lower left corner of the image obtained by MPS were more clear than the other five methods, the brain vein was also clearer, and the contrast was higher than the other five methods. In **Figure 14**, the fused images obtained by LP and SR methods were

TABLE 5 | Comparison of objective evaluation metrics of CT and MRI images.

Figure	Metrics	DWT	LP	SR	DTCWT-SR	VGG	MPS
Figures 12A,B	*SD*	56.2694	60.5508	59.2485	72.1939	49.4054	**73.6843**
	MI	0.6266	0.6713	0.7032	0.7449	0.6849	**1.0761**
	SSIM	0.6846	0.7114	0.7088	0.6631	0.5067	**0.7318**
	Q_G	0.6618	0.6706	0.6818	0.6728	0.3410	**0.7696**
	Q_P	0.3756	0.4625	0.2952	0.3945	0.5255	**0.6941**
Figures 13A,B	*SD*	75.3185	77.7486	80.6627	82.7177	68.0464	**86.4508**
	MI	0.4175	0.4496	0.6142	0.4336	0.5063	**0.7824**
	SSIM	0.5358	0.5861	0.5921	0.5969	0.5787	**0.6041**
	Q_G	0.4343	0.5262	0.5425	0.4216	0.4322	**0.5806**
	Q_P	0.2928	0.4069	0.3859	0.3193	0.3986	**0.5129**
Figures 14A,B	*SD*	45.5362	53.7899	55.2056	53.5713	37.3261	**59.2069**
	MI	0.4510	0.4551	0.8655	0.3547	0.6078	**1.1221**
	SSIM	0.3849	0.4403	0.4937	0.5016	0.3702	**0.5057**
	Q_G	0.6453	0.6430	0.8465	0.8056	0.6750	**0.9191**
	Q_P	0.2833	0.5291	0.5418	0.5613	0.5755	**0.5769**
Figures 15A,B	*SD*	62.0558	66.6555	65.3679	69.7711	55.0330	**72.7695**
	MI	0.6098	0.6060	0.7530	0.4336	0.6993	**0.9532**
	SSIM	0.5962	0.6136	0.6569	0.6410	0.5476	**0.6756**
	Q_G	0.5955	0.5743	0.6552	0.3626	0.3392	**0.7100**
	Q_P	0.2531	0.4005	0.3128	0.2664	0.3675	**0.6454**

Bold values mean maximum value of the same metrices in the same group of comparative experiments.

fractured at the lower right corner. Although DTCWT-SR and VGG methods obtained relatively complete fusion images, there is a certain degree of contrast loss. In **Figure 15**, LP, DTCWT-SR, and VGG-Net methods have some contrast loss, especially in the middle part, at the same time, the image obtained by the SR method presents spatial dislocation at both sides of the eyeball and a certain degree of distortion appears at the position of white connection of the two images. The SR method also has similar shortcomings in this regard, please refer to the lower right corner of the image.

5. CONCLUSION

In this article, we propose a method based on MPS for multi-modal image fusion. First, the source images are initialized into a three-dimensional tensor, and then the tensor is decomposed into several sub-tensors by using a sliding window to obtain the corresponding features. The core matrix is fused by the fusion rule based on the sigmoid function, and the fused image is obtained by multiplying the left-right factor matrix. In this article, we use a sliding window to avoid blocking effects, and fully consider the local information of the source images by dividing the source image into a set of sub-images. The experimental results show that the proposed algorithm is feasible and effective for image fusion. Being different from the average fusion rule of the multi-scale method and the "Max-L1" fusion rule of the SR method, the fusion rule based on the sigmoid function used in the article is more effective, but it also makes the fusion process more complicated of the proposed method. Future study will focus on further exploring a more effective fusion rule to improve the 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/s.

AUTHOR CONTRIBUTIONS

YL designed the algorithm. RW performed experiments and wrote this article. QG, DS, and DZ revised the manuscript. All authors confirmed the submitted version.

FUNDING

This work was supported by Nature Science Foundation of Anhui (2008085MF183), The Key Science Program of Anhui Education Department (KJ2018A0012), The Research Fund for Doctor of Anhui University (J01003266), and National Natural Science Foundation of China (NSFC)(61402004).

ACKNOWLEDGMENTS

The authors would like to thank the members in the lab for discussion and valuable suggestions for this study.

REFERENCES

Bengua, J. A., Phien, H. N., and Tuan, H. D. (2015). "Optimal feature extraction and classification of tensors via matrix product state decomposition," in *2015 IEEE International Congress on Big Data*, 669–672. doi: 10.1109/BigDataCongress.2015.105

Bengua, J. A., Phien, H. N., Tuan, H. D., and Do, M. N. (2017a). Efficient tensor completion for color image and video recovery: low-rank tensor train. *IEEE Trans. Image Process*. 26, 2466–2479. doi: 10.1109/TIP.2017.2672439

Bengua, J. A., Phien, H. N., Tuan, H. D., Do, M. N. (2017b). Matrix product state for higher-order tensor compression and classification. *IEEE Trans. Signal Process*. 65, 4019–4030. doi: 10.1109/TSP.2017.2703882

Burt, P., and Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Trans. Commun*. 31, 532–540. doi: 10.1109/TCOM.1983.1095851

Goshtasby, A., and Nikolov, S. (2007). Image fusion: advances in the state of the art. *Inform. Fusion* 8, 114–118. doi: 10.1016/j.inffus.2006.04.001

Haribabu, M., and Bindu, C. H. (2017). "Visibility based multi modal medical image fusion with DWT," in *IEEE Int. Conf. Power, Control, Signals and Instrum. Eng*., 1561–1566. doi: 10.1109/ICPCSI.2017.8391973

Hong, R. (2000). Objective image fusion performance measure. *Military Tech. Courier* 56, 181–193. doi: 10.5937/vojtehg0802181B

Hui, L., Wu, X. J., and Kittler, J. (2018). "Infrared and visible image fusion using a deep learning framework," in *International Conference on Pattern Recognition 2018*.

Jiang, Y., and Wang, M. (2014). Image fusion with morphological component analysis. *Inform. Fusion* 18, 107–118. doi: 10.1016/j.inffus.2013.06.001

Kolda, T. G., and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev*. 51, 455–500. doi: 10.1137/07070111X

Li, W. S., Du, J., Zhao, Z. M., and Long, J. (2019). Fusion of medical sensors using adaptive cloud model in local Laplacian pyramid domain. *IEEE Trans. Biomed. Eng*. 66, 1172–1183. doi: 10.1109/TBME.2018.2869432

Liang, J. L., He, Y., Liu, D., and Zeng, J. (2012). Image fusion using higher order singular value decomposition. *IEEE Trans. Image Process*. 21, 2898–2909. doi: 10.1109/TIP.2012.2183140

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2016). Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett*. 23, 1882–1886. doi: 10.1109/LSP.2016.2618776

Ma, J., Ma, Y., and Li, C. (2019). Infrared and visible image fusion methods and applications: a survey. *Inform. Fusion* 45, 153–178. doi: 10.1016/j.inffus.2018.02.004

Perez-Garcia, D., Verstraete, F., Wolf, M. M., and Cirac, J. I. (2006). Matrix product state representations. *Physics* 7, 401–430. doi: 10.26421/QIC7.5-6-1

Sanz, M., Egusquiza, I. L., Candia, R. D., Lamata, L., and Solano, E. (2016). Entanglement classification with matrix product states. *Sci. Rep*. 6:30188. doi: 10.1038/srep30188

Schollwock, U. (2011). The density-matrix renormalization group in the age of matrix product states. *Ann. Phys*. 326, 96–192. doi: 10.1016/j.aop.2010.09.012

Schuch, N., Perez-Garcia, D., and Cirac, I. (2011). Classifying quantum phases using matrix product states and projected entangled pair states. *Phys. Rev*. 84, 667–673. doi: 10.1103/PhysRevB.84.165139

Singh, R., Srivastava, R., Prakash, O., and Khare, A. (2012). "DTCWT based multimodal medical image fusion," in *International conference on Signal, Image and Video processing (ICSIVP-2012)*, 403–407.

Yang, B., and Li, S. (2010). Multifocus image fusion and restoration with sparse representation. *IEEE Trans. Instrument. Measure*. 59, 884–892. doi: 10.1109/TIM.2009.2026612

Yu, P., Sun, Q. S., and Xia, D. (2011). Image fusion framework based on decomposition of PCA. *Comput. Eng*. 37, 210–213.

Zhang, Q., and Levine, M. D. (2016). Robust multi-focus image fusion using multi-task sparse representation and spatial context. *IEEE Trans. Image Process*. 25, 2045–2058. doi: 10.1109/TIP.2016.2524212

Zhang, X., Wen, G., and Dai, W. (2016). A tensor decomposition-based anomaly detection algorithm for hyperspectral image. *IEEE Trans. Geosci. Remote Sens*. 54, 5801–5820. doi: 10.1109/TGRS.2016.2572400

Zhang, Y. (2004). Understanding image fusion. *Photogrammetr. Eng. Remote Sens*. 70, 657–661.

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.

frontiers
in Neurorobotics

ORIGINAL RESEARCH
published: 25 November 2021
doi: 10.3389/fnbot.2021.775688

Boosting-GNN: Boosting Algorithm for Graph Networks on Imbalanced Node Classification

*Shuhao Shi, Kai Qiao, Shuai Yang, Linyuan Wang, Jian Chen and Bin Yan**

Henan Key Laboratory of Imaging and Intelligence Processing, PLA Strategy Support Force Information Engineering University, Zhengzhou, China

The graph neural network (GNN) has been widely used for graph data representation. However, the existing researches only consider the ideal balanced dataset, and the imbalanced dataset is rarely considered. Traditional methods such as resampling, reweighting, and synthetic samples that deal with imbalanced datasets are no longer applicable in GNN. This study proposes an ensemble model called Boosting-GNN, which uses GNNs as the base classifiers during boosting. In Boosting-GNN, higher weights are set for the training samples that are not correctly classified by the previous classifiers, thus achieving higher classification accuracy and better reliability. Besides, transfer learning is used to reduce computational cost and increase fitting ability. Experimental results indicate that the proposed Boosting-GNN model achieves better performance than graph convolutional network (GCN), GraphSAGE, graph attention network (GAT), simplifying graph convolutional networks (SGC), multi-scale graph convolution networks (N-GCN), and most advanced reweighting and resampling methods on synthetic imbalanced datasets, with an average performance improvement of 4.5%.

Keywords: graph neural network, imbalanced datasets, ensemble learning, adaboost, node classification

OPEN ACCESS

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Hao Peng,
Beihang University, China
Wei Huang,
University of Electronic Science and Technology of China, China

****Correspondence:***
Bin Yan
ybspace@hotmail.com

Received: *14 September 2021*
Accepted: *21 October 2021*
Published: *25 November 2021*

Citation:
Shi S, Qiao K, Yang S, Wang L, Chen J and Yan B (2021) Boosting-GNN: Boosting Algorithm for Graph Networks on Imbalanced Node Classification. Front. Neurorobot. 15:775688. doi: 10.3389/fnbot.2021.775688

1. INTRODUCTION

Convolutional neural networks (CNNs) have been widely used in image recognition (Russakovsky et al., 2015; He et al., 2016), object detection (Lin et al., 2014), speech recognition (Yu et al., 2016), visual coding and decoding (Huang et al., 2021a,b). However, traditional CNNs can only handle data in the Euclidean space. It cannot effectively address graphs that are prevalent in real life. Graph neural networks (GNNs) can effectively construct deep learning models on graphs. In addition to homogeneous graphs, heterogeneous GNN (Wang et al., 2019; Li et al., 2021; Peng et al., 2021) can effectively handle more comprehensive information and semantically richer heterogeneous graphs.

The graph convolutional network (GCN) (Kipf and Welling, 2016) has achieved remarkable success in multiple graph data-related tasks, including recommendation systems (Chen et al., 2020; Yu and Qin, 2020), molecular recognition (Zitnik and Leskovec, 2017), traffic forecast (Bai et al., 2020), and point cloud segmentation (Li et al., 2019). GCN is based on the neighborhood aggregation scheme, which generates node embedding by combining information from neighborhoods. GCN achieves superior performance in solving node classification problems compared with conventional methods, but it is adversely affected by datasets imbalance. However, existing studies on GCNs all aim at balanced datasets, and the problem of imbalanced datasets have not been considered.

In the field of machine learning, the processing of imbalanced data sets is always challenging (Carlson et al., 2010; Taherkhani et al., 2020). The data distribution of an imbalanced dataset makes the fitting ability of the model insufficient because it is difficult for the model to learn useful information from unevenly distributed datasets (Japkowicz and Stephen, 2002). A balanced dataset consists of almost the same number of training samples in each class. In reality, it is impractical to obtain the same number of training samples for different classes because the data in different classes are generally not uniformly distributed (Japkowicz and Stephen, 2002; Han et al., 2005). The imbalance of the training dataset is caused by many possible factors, such as deviation sampling and measurement errors. Samples may be collected from narrow geographical areas in a specific time period and in different areas at different times, exhibiting a completely different sample distribution. The datasets widely used in deep learning research, e.g., IMAGENET large scale visual recognition challenge (ImageNet ILSVRC 2012) (Russakovsky et al., 2015), microsoft common objects in context (MS COCO) (Lin et al., 2014), and Places Database (Zhou et al., 2018), balanced datasets, where the amount of data in different classes is basically the same. Recently, more and more imbalanced datasets reflecting real-world data characteristics have been built and released, e.g., iNaturalist (Cui et al., 2018), a dataset for large vocabulary instance segmentation (LVIS) (Gupta et al., 2019), and a large-scale retail product checkout dataset (RPC) (Wei et al., 2019). It is difficult for traditional pattern recognition methods to achieve excellent results on imbalanced datasets, so methods that can deal with imbalanced datasets efficiently are urgently needed.

For imbalanced datasets, additional processing is needed to reduce the adverse effects (Japkowicz and Stephen, 2002). The existing machine learning methods mainly rely on resampling, data synthesis, and reweighting. 1) Resampling samples the original data by undersampling and oversampling. Undersampling removes part of data in the majority class so that the majority class can match with the minority class in terms of the amount of data. Oversampling copies the data in the minority class. 2) Data synthesis, i.e., synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) and its improved version (Han et al., 2005; Ramentol et al., 2011; Douzas and Bação, 2019) as well as other synthesis methods (He et al., 2008), synthesize the new sample artificially by analyzing the samples in the minority class. 3) Reweighting assigns different weights to different samples in the loss function to improve the model's performance of the model on imbalanced datasets.

In the GNN, the existing processing methods for imbalanced datasets in machine learning are not applicable. 1) The data distribution problem of imbalanced datasets cannot be overcome by resampling. The use of oversampling may introduce many repeated samples to the model, which reduces the training speed and leads to overfitting easily. In the case of undersampling, valuable samples that are important to feature learning may be discarded, making it difficult for the model to learn the actual data distribution. 2) The use of the data synthesis method or oversampling method loses the relationship between the newly generated samples and the original samples in the GNN, which affects the aggregation process of nodes. 3) Reweighting, e.g., Focal Loss (Lin et al., 2017), and CB Focal Loss (Cui et al., 2019), can solve the problem of the imbalanced dataset in GCN to some extent, but it does not consider the relationship between training samples, and fails to achieve satisfactory performance in dealing with imbalanced datasets.

Ensemble learning methods are more effective in improving the classification performance of imbalanced data than data sampling techniques (Khoshgoftaar et al., 2015). It is challenging for a single model to classify rare and few samples on an imbalanced dataset accurately, thus, the overall performance is limited. Ensemble learning is a process of aggregating multiple base classifiers to improve the generalization ability of classifiers. Briefly, ensemble learning uses multiple weak classifiers to make classification on the dataset. In traditional machine learning, ensemble learning is used to improve the classification accuracy of multi-class imbalanced data (Chawla et al., 2003; Seiffert et al., 2010; Galar et al., 2013; Blaszczynski and Stefanowski, 2015; Nanni et al., 2015; Hai-xiang et al., 2016). In CNNs, some models adopt ensemble learning to deal with imbalanced datasets. Enhanced-random-feature-subspace-based ensemble CNN (Lv et al., 2021) adaptively resamples the training set in iterations to get multiple classifiers and forms a cascade ensemble model. AdaBoost-CNN (Taherkhani et al., 2020) integrates AdaBoost with a CNN to improve accuracy on imbalanced data.

Inspired by ensemble learning, an ensemble GNN classifier that can deal with the imbalanced dataset is proposed in this study. The adaptive boosting (AdaBoost) algorithm is combined with GNN to train the GNN classifier by serialization, and the samples are reweighted according to the calculation results. Based on this, the proposed classifier improved the classification performance on the imbalanced dataset. The main contributions of this study are as follows:

- This article uses the ensemble learning to study the imbalanced dataset problem in GNN for the first time. An Boosting-GNN model is proposed to deal with imbalanced datasets in semi-supervised nodes classification. A transfer learning strategy is also applied to speed up the training of the Boosting-GNN model.
- Four imbalanced datasets are constructed to evaluate the performance of the Boosting-GNN. Boosting-GNN uses GCN, GAT, and GraphSAGE as base classifiers, improving the classification accuracy on imbalanced datasets.
- The robustness of Boosting-GNN under feature noise perturbations is discussed, and it is discovered that ensemble learning can significantly improve the robustness of GNNs.

The rest of this article is organized as follows. Section 2 introduces the related work of dealing with imbalanced data sets and the application of ensemble learning in deep learning. In section 3, the principle of the proposed Boosting-GNN is discussed. Then, four datasets and a proposed method for performance evaluation are described, and the experimental results are discussed in section 4. Finally, section 5 concludes the article.

2. RELATED WORKS

Due to the prevalence of imbalanced data in practical applications, the problem of imbalanced data sets has attracted more and more attention. Recent researches are mainly conducted in the following four directions:

2.1. Resampling

Resampling can be specifically divided into two types: 1) Oversampling by copying data in minority classes (Buda et al., 2018; Byrd and Lipton, 2019). After oversampling, some samples are repeated in the dataset, leading to a less robust model and worse generalization performance on imbalanced data. 2) Undersampling by selecting data in the majority classes (Buda et al., 2018; Byrd and Lipton, 2019). Undersampling may cause information loss in majority classes. The model only learns a part of the overall pattern, leading to underfitting (Shen and Lin, 2016). K-means and stratified random sampling (KSS) (Zhou et al., 2020) performs undersampling after K-means clustering for majority classes, and achieves good results.

2.2. Synthetic Samples

The data synthesis methods generate samples similar to samples of minority classes in the original set. The representative method is SMOTE (Chawla et al., 2002), and the operations of this method are as follows. For each sample in a small sample set, an arbitrary sample is selected from its K-nearest neighbors. Then, a random point on the line between the sample and the selected sample is taken as a new sample. However, the overlapping degree will be increased by synthesizing the same number of new samples for each minority class. The Borderline-SMOTE (Han et al., 2005) synthesizes new samples similar to the samples on the classification boundary. Preprocessing method combining SMOTE and RST (SMOTE-RSB*) (Ramentol et al., 2011) exploits the synthetic minority oversampling technique and the editing technique based on the rough set theory. Geometric SMOTE (G-SMOTE) (Douzas and Bação, 2019) generates a synthesized sample for each of the selected instances in a geometric region of the input space. Adaptive synthetic sampling (ADASYN) (He et al., 2008) algorithm synthesizes different number of new samples for different minority classes samples.

2.3. Reweighting

Reweighting typically assigns different weights to different samples in the loss function. In general, reweighting assigns large weights to training samples in minority classes (Wang et al., 2017). Besides, finer control of loss can be achieved at the sample level. For example, Focal Loss (Lin et al., 2017) designed a weight adjustment scheme to improve the classification performance of imbalanced dataset. CB Focal Loss (Cui et al., 2019) introduced a weight factor inversely proportional to the number of effective samples to rebalance the loss, reaching the most advanced level in the imbalanced dataset.

2.4. Ensemble Classifiers

Ensemble classifiers are more effective than sampling methods to deal with the imbalance problem (Khoshgoftaar et al., 2015). In GNN models, AdaGCN (Sun et al., 2021) integrates Adaboost and GCN layers to get deeper network models. Different from AdaGCN, Boosting-GNN uses GNN as a sub-classifier of Boosting algorithm to improve the performance on imbalanced datasets. To our knowledge, we are the first to use ensemble learning to solve the classification on graph imbalanced datasets.

In addition, there are transfer learning, domain adaptation, and other methods to deal with imbalance problems. The method based on transfer learning solves the problem by transferring the characteristics learned from majority classes to minority classes (Yin et al., 2019). Domain adaptive method processes different types of data and learns how to reweight adaptively (Zou et al., 2018). These methods are beyond the scope of this article.

3. THE PROPOSED METHOD

3.1. GCN Model

Given an input undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$, respectively, denote the set of N nodes and the set of e edges. The corresponding adjacency matrix $A \in \mathbb{R}^{N \times N}$ is an $N \times N$ sparse matrix. The entry (i, j) in the adjacency matrix is equal to 1 if there is an edge between i and j, and 0, otherwise. The degree matrix D is a diagonal matrix where each entry on the diagonal indicates the degree of a vertex, which can be computed as $d_i = \sum_j a_{ij}$. Each node is associated with an F-dimensional feature vector, and $X \in \mathbb{R}^{N \times F}$ denotes the feature matrix for all nodes. GCN model of semi-supervised classification has two layers (Kipf and Welling, 2016), and every layer computes the transformation:

$$H^{(l+1)} = \sigma(Z^{(l+1)}), Z^{(l+1)} = \tilde{A}H^{(l)}W^{(l)} \tag{1}$$

where $\tilde{A}$ is normalized adjacency obtained by $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. $W^{(l)}$ is the trainable weights of the layer. $\sigma(\cdot)$ denotes an activation function (usually ReLU), and $H^{(l)} \in \mathbb{R}^{N \times d_l}$ is the input activation matrix of the łth hidden layer, where each row is a d_l-dimensional vector for node representation. The initial node representations are the original input features:

$$H^{(0)} = X \tag{2}$$

A two-layer GCN model can be defined in terms of vertex features X and $\hat{A}$ as:

$$\mathrm{GCN}_{2-\mathrm{layer}}(\hat{A}, X; \theta) = softmax(\hat{A} \cdot \sigma(\hat{A}XW^{(0)})W^{(1)}) \tag{3}$$

The GCN is trained by the back propagation learning algorithm. The last layer uses the *softmax* function for classification, the cross-entropy loss over all labeled examples are evaluated:

$$\mathcal{L} = -\sum_{|\mathcal{Y}_L|}\sum_{i \in \mathcal{Y}_L} loss(y_i, z_i^L) \tag{4}$$

Formally, given a dataset with n entities $(X, Y) = \left\{(x_i, y_i)\right\}_{i=1}^{N}$, where x_i represents the word embedding for entity i, and $y_i \in \{1, \cdots\cdots, C\}$ represents the label for x_i. Multiple weak classifiers are combined with AdaBoost algorithm to make a single strong classifier.

3.2. Proposed Algorithm

Since ensemble learning is an effective method to deal with imbalanced datasets, Boosting-GNN adopts the Adaboost algorithm proposed by Hastie et al. (2009) to design an ensemble strategy for GCNs, which can train the GCNs sequentially. In Boosting-GNN, the weight of each training sample is assigned according to the degree to which the sample was not correctly trained in the previous classifier.

3.2.1. Aggregation

Boosting-GNN aggregates GNN through the Adaboost algorithm to improve the performance on imbalanced datasets. First, the overall formula of Boosting-GNN can be expressed as:

$$F_M(x) = \sum_{m=1}^{M} \alpha_m * G_m(x; \theta_m) \tag{5}$$

where $F_M(x)$ is the ensemble classifier obtained after M rounds of training, and x denotes samples. A new GNN classifier $G_m(x; \theta_m)$ is trained in each round, and θ_m is the optimal parameter learned by the base classifier. The weight of the classifier α_m denotes the importance of classifier, and it could be obtained according to the error of the classifier. Based on (5), Formula (6) can be obtained:

$$F_m(x) = F_{m-1}(x) + \alpha_m * G_m(x; \theta_m) \tag{6}$$

$F_{m-1}(x)$ is the weighted aggregation of the previously trained base classifier. In each iteration, a new base classifier $G_m(x; \theta_m)$ and its weights α_m are solved. Boosting-GNN uses an exponential loss function:

$$L(y, F(x)) = e^{-y*F(x)} \tag{7}$$

According to the meaning of the loss function, if the classification is correct, the exponent part is a negative number, otherwise, it is a positive number. As for training the base classifier, the training dataset is $T = \left\{(x_i, y_i)_{i=1}^{N}\right\}$, x_i is the feature vector of the ith node; y_i is the category label of the ith node, and $y_i \in \{1, \ldots, C\}$, where C is the total number of classes.

3.2.2. Reweight Samples

Assume that during the first training, the samples are evenly distributed and all weights are the same. The data weights are initialized by $D_1 = \left\{w_1^1, w_2^1, \ldots, w_N^1\right\}$, where $w_i^1 = 1/N, i = 1, \ldots, N$, and N is the number of samples. Training M networks in sequence on the training set, the expected loss ε_m at the mth iteration is:

$$\varepsilon_m = \sum_{y_i \neq G_m(x_i;\theta_m)} w_i^m = \sum_{i=1}^{N} w_i^m \mathbb{I}(y_i \neq G_m(x_i; \theta_m)) \tag{8}$$

where $\mathbb{I}$ is the indicator function. When the input is true, the function value is 1; otherwise, the function value is 0. ε_m is the sum of the weights of all misclassified samples. α_m can be treated as a hyper-parameter to be tuned manually, or as a model parameter to be optimized automatically. In our model, to keep it simple, α_m is assigned according to ε_m.

$$\alpha_m = \frac{1}{2} \ln \frac{1 - \varepsilon_m}{\varepsilon_m} \tag{9}$$

α_m decreases as ε_m increases. The first GNN is trained on all the training samples with the same weight of $1/N$, indicating the same importance for all samples. After the M estimators are trained, the output of GNN can be obtained, which is a C-dimensional vector. The vector contains the predicted values of C classes, which indicate the confidence of belonging to the corresponding class. For the mth GNN input sample x_i, the output vector is $p^m(x_i)$. $p_k^m(x_i)$ is the kth element of $p^m(x_i)$, where $k = 1, 2, \cdot \cdot \cdot, C$.

$$w_i^{m+1} = w_i^m e^{\left(-a \frac{C-1}{C} y_i \log\left(p^m(x_i)\right)\right)} \tag{10}$$

w_i^m is the weight of the ith training sample of the mth GNN. y_i is the one-hot label vector encoded according to the ith training sample. Formula (10) is obtained based on Adaboost's Samme.r algorithm (Hastie et al., 2009), which is used to update the weight of the sample. If the output vector of the misclassified sample is not related to the output label, a large value is obtained for the exponential term, and the misclassified sample will be assigned a larger weight in the next GNN classifier. Similarly, a correctly classified sample will be assigned a smaller weight in the next GNN classifier. In summary, the weight vector D is updated so that the weight of the correctly classified samples is reduced and the weight of the misclassified samples is increased.

After the weights of all training samples for the current GNN are updated, they are normalized by the sum of weights of all samples. When the classifier $F_m(x)$ is trained, the weight distribution of the training dataset is updated for the next iteration. When the subsequent GNN-based classifier is trained, the GNN training does not start from a random initial condition. Instead, the parameters learned from the previous GNN are transferred to the $(m + 1)$th GNN, so GNN is fine-tuned based on the previous GNN parameters. The use of transfer learning can reduced the number of training epochs and make the model fit faster.

Moreover, due to the change of weight, the subsequent GNN focuses on untrained samples. The subsequent GNN performs training from scratch on a small number of training samples, which easily causes overfit. For a large number of training samples, the expected label output $p^m(x_i)$ by the GNN after training has a strong correlation with the real label y_i. For the subsequent GNN classifier, the trained samples have a smaller weight than the sample without previous GNN training.

3.2.3. Testing With Boosting-GNN

After training the M base classifiers, Equation (11) can be used to predict the category of the input sample. The outputs of M base classifiers are summed. In the summed probability vector, the category with the highest confidence is regarded as the predicted category.

$$Q(x) = \underset{k}{argmax} \sum_{m=1}^{M} h_k^m(x) \tag{11}$$

h_k^m is the classification result of the kth sample made by the mth basis classifier, $k = 1, 2, \cdot \cdot \cdot, C$, which can be calculated from the

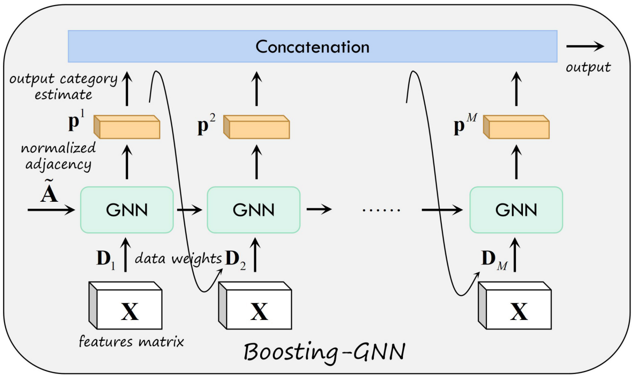


FIGURE 1 | Schematic of the proposed Boosting-GNN.

Equation (12).

$$h_k^m = (C-1)\cdot\left(\log\left(p_k^m(x)\right) - \frac{1}{C}\sum_{i=1}^{C}\log\left(p_i^m(x)\right)\right) \qquad (12)$$

Where $p_i^m(x)$ is the kth element of the output vector of the mth GCN classifier for the input x. **Figure 1** shows the schematic of the proposed Boosting-GNN. The first GNN is first trained with the initial weight D_1. Then, based on the output of the first GNN, the data weight D_2 used to update the second GNN are obtained. In addition, the parameters learned from the first GNN are transferred to the second GNN. After the mth base classifier is trained in order, all base classifiers are aggregated to obtain the final Boosting-GNN classifier.

The pseudo-code for an Boosting-GNN is exhibited in **Algorithm 1**. In each iteration of sequential learning, the classifiers are first trained with corresponding training data and weights. Then, according to the training results of the classifiers, the data weights are updated for the next iteration. Both operations are performed until M base classifiers are trained.

Algorithm 1 Framework of the Boosting-GNN algorithm.

```
Input: Training set $T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$;
Output: Ensemble of classifiers $F_M(x)$;
Initialization: $w_i^1 = 1/N$ for all $1 \le i \le N$
for $m = 1, 2, \cdots, N$ do;
    if $m = 1$ then
        Train GNN classifier with weighted sample set $\{T, D_1\}$;
    else
        Transfer the learning parameters of the $(m-1)$th GNN to the $m$th GNN classifier;
        Train the $m$th GNN classifier with weighted sample set;
    end if
    Calculate the output category estimated for the $C$ classes of the $m$th GNN classifier $p_k^m(x)$, where $k = 1, 2, \cdots, C$;
    Calculate the training error $\varepsilon_m$ of the $m$th classifier according to (8);
    Assign the weight $\alpha_m$ to the classifier based on $\varepsilon_m$ using (9);
    Update the sample weight $D_{m+1}$ according to $p_k^m(x)$, and normalize the sample weight $D_{m+1}$;
end for
```

4. EXPERIMENTS AND ANALYSIS

4.1. Experimental Settings

The proposed ensemble model is evaluated on three well-known citation network datasets prepared by Kipf and Welling (2016): Cora, Citeseer, and Pubmed (Sen et al., 2008). These datasets are chosen because they are available online and are used by our baselines. In addition, experiments are also conducted on the Never-Ending Language Learning (NELL) dataset (Carlson et al., 2010). As a bipartite graph dataset extracted from a knowledge graph, NELL has a larger scale than the citation datasets, and it has 210 node classes.

4.1.1. Citation Networks

The nodes in the citation datasets represent articles in different fields, and the labels of nodes represent the corresponding journal where the articles were published. The edges between two nodes represent the reference relationship between articles. If an edge exists between the nodes, there is a reference relationship between the articles. Each node has a one-hot vector corresponding to the keywords of the article. The task of categorization is to classify the domain of unlabeled articles based on a subset of tagged nodes and references to all articles.

TABLE 1 | Datasets used for experiments.

Dataset	Cora	Citeseer	Pubmed	NELL
Vertices	2,708	3,327	19,717	65,755
Edges	5,429	4,732	44,338	266,144
Classes	7	6	3	210
Features	1,433	3,703	500	5,414

4.1.2. Never-Ending Language Learning

The pre-processing schemes described in Yang et al. (2016) are adopted in this study. Each relationship is represented as a triplet (e_1, r, e_2), where e_1, r, and e_2, respectively, represent the head entity, the relationship, and the tail entity. Each entity E is regraded as a node in the graph, and each relationship r consists of two nodes r_1 and r_2 in the graph. For each (e_1, r, e_2), two edges (e_1, r_1) and (e_2, r_2) are added to the graph. A binary, symmetric adjacency matrix from this graph is constructed by setting entries $A_{ij} = 1$, if one or more edges are present between nodes i and j (Kipf and Welling, 2016). All entity nodes are described by sparse feature vectors with the dimension of 5,414. **Table 1** summarizes the statistics of these datasets.

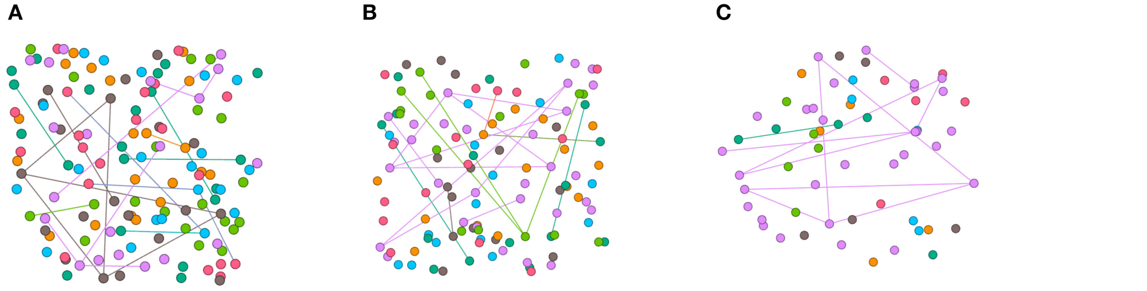


FIGURE 2 | Visualization of synthetic imbalanced datasets. **(A)** shows the classical Cora training set. **(B)** shows the training set when *s* is fetched 15. **(C)** shows the training set when *s* is fetched 5. The mean degrees of the nodes in **(A–C)** are 0.30, 0.30, 0.37 respectively.

4.1.3. Synthetic Imbalanced Datasets

Different synthetic imbalanced datasets are constructed based on the datasets mentioned above. According to the Pareto Principle that 80% of the consequences come from 20% of the causes, one of the classes is randomly selected as the majority category for simplicity. The remaining classes are regraded as minority classes. In Kipf and Welling (2016), 20 samples of each class were selected as the training set, and to keep the number of training samples broadly consistent, the datasets are described in Equation (13).

$$n_i = \begin{cases} 30 & i = c \\ s & i \neq c \end{cases} \tag{13}$$

n_i is the number of samples in category i, c is the randomly selected category, C is the number of classes in the dataset, and s is the number of samples in the minority category. By changing s, the number of minority category samples is altered, thus changing the degree of imbalance in the training set. For example, in the Cora dataset, there are seven classes of samples. So, the number of samples in one class is fixed to 30, and the number of samples in the other six classes is changed. Each time the training is conducted, a certain number of samples are randomly selected to form the training set. The test set is divided following the method in Kipf and Welling (2016) to evaluate the performance of different models.

Synthetic imbalanced datasets are constructed by node dropping. Given the graph $\mathcal{G}$, node dropping will randomly discard vertices along with their connections until the number of different classes of nodes matches the setting. In node dropping, the dropping probability of each node follows a uniform distribution. We visualize the synthetic datasets in **Figure 2** and use different colors to represent different categories of nodes. Due to the sparsity of the adjacency matrix of the graph data set, imbalanced sampling of the graph data does not reduce the average degree of the nodes. Although disconnect parts of the graph, missing part of vertices does not affect the semantic meaning of $\mathcal{G}$.

4.1.4. Parameter Settings

In Boosting-GNN, five GNN base classifiers are used. Boosting-GNN, respectively, uses GCN, GraphSAGE, and GAT as the base classifiers. All networks are composed of two layers, and all models are trained for a maximum of 100 epochs (training iterations) using Adam optimizer. For Cora, Citeseer, and Pubmed datasets, the number of hidden units is 16, and L2 regularization is 5e-4. For NELL, the number of hidden units is 128, and L2 regularization is 1e-5.

The following sets of hyperparameters are used for Boosting-GNN: For Boosting-GCN, the activation function is ReLU. The learning rates on Cora, Citeseer, Pubmed, and NELL are 1e-2, 1e-2, 1e-2, 5e-3, respectively. For Boosting-GraphSAGE, the activation function is ReLU. The sampled sizes (S1 = 25, S2 = 10) is used for each layer. The learning rates on Cora, Citeseer, Pubmed, and NELL are 1e-3, 1e-3, 5e-4, 1e-4, respectively. For Boosting-GAT, the first-layer activation function is *ELU* and the second-layer activation function is *softmax*. The number of attention heads K is 8. The learning rates on Cora, Citeseer, Pubmed and NELL are 1e-3, 1e-3, 1e-3, 5e-4, respectively.

For GCN, GraphSAGE, GAT, SGC, N-GCN, and other algorithms, the models are trained for a total of 500 epochs. The highest accuracy is taken as the result of a single experiment, and the mean accuracy of 10 runs with random sample split initializations is taken as the final result. A different random seed is used for every run (i.e., removing different nodes), but the 10 random seeds are the same across models. All the experiments are conducted on a machine equipped with two NVIDIA Tesla V100 GPU (32 GB memory), 20-core Intel Xeon CPU (2.20 GHz), and 192 GB of RAM.

4.2. Baseline Methods

The performance of the proposed method is evaluated and compared to that of three groups of methods:

4.2.1. GCN Methods

In experiments, our Boosting-GNN model is compared with the following representative baselines:

- Graph convolutional network (Kipf and Welling, 2016) produces node embedding vectors by truncating the Chebyshev polynomial to the first-order neighborhoods.
- GAT (Velickovic et al., 2018) generates node embedding vectors for each node by introducing an attention mechanism when computing node and its neighboring nodes.

TABLE 2 | Summary of results in terms of classification accuracy (in percentage).

Model	Cora	Citeseer	Pubmed	NELL
GCN	65.6 ± 0.8	62.2 ± 0.5	71.8 ± 0.6	68.5 ± 1.4
GraphSAGE	66.3 ± 0.8	59.7 ± 0.6	69.7 ± 0.6	69.6 ± 1.3
GAT	67.4 ± 0.7	60.3 ± 0.6	66.2 ± 0.7	70.3 ± 1.6
N-GCN	67.3 ± 0.6	65.4 ± 0.3	72.3 ± 0.3	73.3 ± 1.2
SGC	69.7 ± 0.8	59.4 ± 0.5	66.9 ± 0.5	67.1 ± 1.4
GCN-FL	67.8 ± 1.2	65.1 ± 0.8	72.4 ± 0.8	71.2 ± 1.2
GraphSAGE-FL	66.5 ± 1.2	59.5 ± 0.8	69.7 ± 1.3	72.1 ± 1.1
GAT-FL	67.4 ± 1.3	61.3 ± 0.7	69.2 ± 1.2	72.6 ± 1.0
GCN-CB	70.6 ± 0.9	65.1 ± 0.6	72.3 ± 0.8	72.9 ± 1.4
GraphSAGE-CB	66.3 ± 0.9	59.7 ± 0.9	70.1 ± 0.9	69.8 ± 1.4
GAT-CB	67.6 ± 1.0	60.3 ± 1.0	69.3 ± 0.9	73.4 ± 1.5
GCN-RS	70.4 ± 1.0	61.8 ± 1.1	70.4 ± 1.1	68.9 ± 2.1
Boosting-GCN	73.2 ± 0.7	**65.7 ± 0.7**	**73.1 ± 0.7**	74.9 ± 1.0
Boosting-GraphSAGE	72.4 ± 1.0	63.2 ± 1.0	70.4 ± 1.1	75.3 ± 1.2
Boosting-GAT	**73.5 ± 0.5**	64.3 ± 0.8	69.7 ± 0.7	**75.5 ± 1.0**

The highest performance of models is highlighted in boldface.

- GraphSAGE (Hamilton et al., 2017) generates the embedding vector of the target vertex by learning a function that aggregates neighboring vertices. The default settings of sampled sizes (S1 = 25, S2 = 10) is used for each layer in GraphSAGE.
- SGC (Wu et al., 2019) reduces model complexity by eliminating the non-linearity between GCN layers, transforming a non-linear GCN into a simple linear model that is more efficient than GCNs and other GNN models for many tasks.
- N-GCN (Abu-El-Haija et al., 2019) obtains the feature representation of nodes by convolving in the neighborhood of nodes at different scales and then fusing all the convolution results. These methods can be regarded as ensemble models.

4.2.2. Resampling Method

The KSS (Zhou et al., 2020) method is used for performance comparison. KSS is a kind of K-means clustering method based on undersampling and achieves state-of-the-art performance on an imbalanced medical dataset.

4.2.3. Reweighting Method

Boosting-GNN is compared with GCN, GraphSAGE, and GAT. These classic models use Focal Loss (Lin et al., 2017) and CB-Focal (Cui et al., 2019), and achieve good classification accuracy on imbalanced datasets.

4.3. Node Classification Accuracy

Our method is implemented in Keras. For the other methods, the code from the Github pages introduced in the original articles is used. For synthetic imbalanced datasets, s is set to 10. The classification accuracy of GCN, GraphSAGE, GAT, SGC, N-GCN, and Boosting-GNN method is listed in **Table 2**.

Results in **Table 2** show that Boosting-GNN outperforms the classic GNN models and state-of-the-art methods for processing imbalanced datasets. The N-GCN obtains a feature representation of the nodes by convolving around the nodes at different scales and then fusing all the convolution results, which can slightly improve the classification compared to the GCN. Resampling method and Reweighting method can improve the accuracy of GNN on imbalanced datasets, but the improvement is very limited. Since RS is not suitable for graph data, RE is slightly better than RS. Boosting-GNN can significantly improve the classification accuracy of GNN, with an average increase of 6.6, 3.7, 1.8, and 5.8% compared with the original GNN model in Cora, Citeseer, Pubmed, and NELL datasets, respectively.

Implementation details are as follows: Following the method in Kipf and Welling (2016), 500 nodes are used as the validation set and 1,000 nodes as the test set. Besides, for a fair performance comparison, the same training procedure is used for all the models.

4.4. Effect of Different Levels of Imbalance in the Training Data

The level of imbalance in the training data is changed by gradually increasing s from 1 to 10. The evaluation results of Boosting-GNN, GCN, GraphSAGE, and GAT are compared, which are shown in **Figure 3**.

Results in **Figure 3** show that classification accuracy of different models varies with s. The shadows indicate the range of fluctuations in the experimental results. When s is relatively small, the degree of imbalance in the training data is large. In this case, the classification accuracy of Boosting-GNN is higher than that of GCN, GraphSAGE, and GAT. As s decreases, the performance advantage of Boosting-GNN increases gradually. Experimental results show that when the sample imbalance is large, aggregation can significantly reduce the adverse effects caused by sample imbalance and improve the classification accuracy. On the Cora dataset, the accuracy of Boosting-GCN, Boosting-GraphSAGE, Boosting-GAT exceeds that of GCN, GraphSAGE, and GAT by 10.3, 8.0, and 6.1% respectively at most.

4.5. Impact of Numbers of Base Classifiers

The number of base classifiers is changed to evaluate the classification accuracy on imbalanced datasets with different base classifiers. We compare the classification results of Boosting-GCN and GCN, and the experimental results are listed in **Table 3**.

The experimental results show that aggregation can contribute to performance improvements. As the number of base classifiers increases, the performance improvement is more and more significant. As the number of base classifiers increases from 3 to 11, the number of base classifiers is odd. The data of Cora, Pubmed, and Citeseer are verified, and the division of train set and test set is the same as that of Section 4.3. Ten experiments are conducted, and each base classifier are trained with 100 epochs and 200

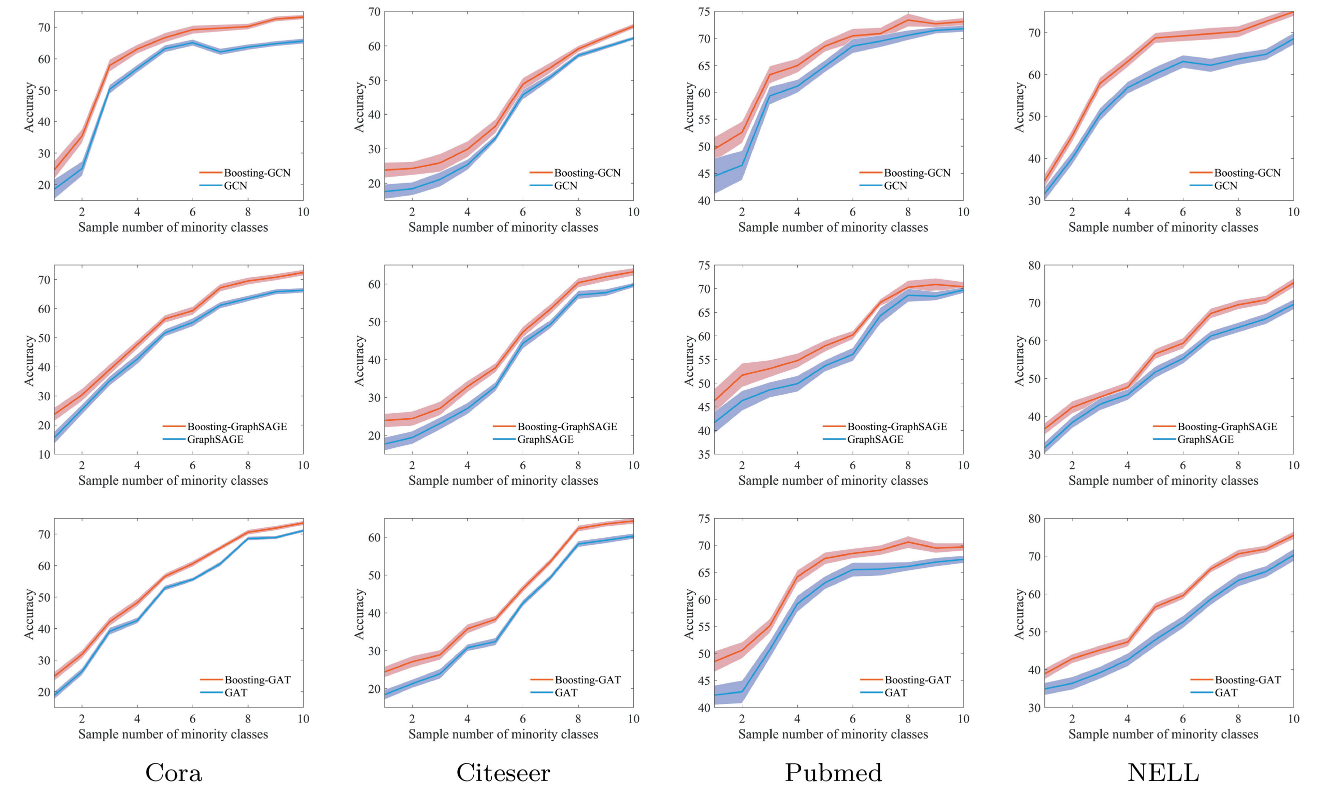


FIGURE 3 | The classification accuracy of Boosting-GNN, graph convolutional network (GCN), GraphSAGE, and GAT on imbalanced datasets.

TABLE 3 | Results of Boosting-GCN with varying numbers of base classifiers in terms of accuracy (in percentage).

Numbers of base classifiers	epoch:100			epoch:200		
	Cora	Citeseer	Pubmed	Cora	Citeseer	Pubmed
3	**75.7 ± 2.4**	65.5 ± 2.5	63.9 ± 2.4	75.4 ± 2.1	65.6 ± 1.1	72.0 ± 0.8
5	73.2 ± 0.7	**65.7 ± 0.7**	73.1±0.7	**75.6 ± 2.3**	**65.9 ± 0.5**	73.1 ± 1.1
7	73.5 ± 1.4	64.5 ± 0.5	**73.5 ± 1.4**	74.1 ± 2.7	64.7 ± 0.4	**73.5 ± 0.8**
9	72.0 ± 0.5	63.6 ± 0.5	72.0 ± 0.5	73.9 ± 2.0	64.2 ± 0.3	72.6 ± 1.1
11	73.0 ± 0.7	64.5 ± 0.6	73.0 ± 0.7	74.1 ± 2.3	65.1 ± 0.3	71.5 ± 0.7

The highest performance of models is highlighted in boldface.

epochs. The training samples are randomly selected for each experiment.

To sum up, when the number of base classifiers is small, the classification accuracy increases with the number of base classifiers. When the number of base classifiers reaches a certain degree, the accuracy decreases due to overfitting.

4.6. Tolerance to Feature Noise

The proposed method is tested under feature noise perturbations by removing node features randomly (Abu-El-Haija et al., 2019). This test is practical, because, in the Citation networks datasets, features could be missing as the authors article might forget to include relevant terms in the article abstract. By removing different features from a node, the performance of Boosting-GNN, GCN, GraghSAGE, and GAT is compared.

Figure 4 shows the performance of different methods when features are removed. As the number of removed features is increased, Boosting-GNN achieves better performance than GCN, GraghSAGE, and GAT. The greater the proportion of features removed, the greater the performance advantage of Boosting-GNN over other models. This suggests that our approach can restore the deleted features to some extent by pulling in the features directly from nearby and distant neighbors.

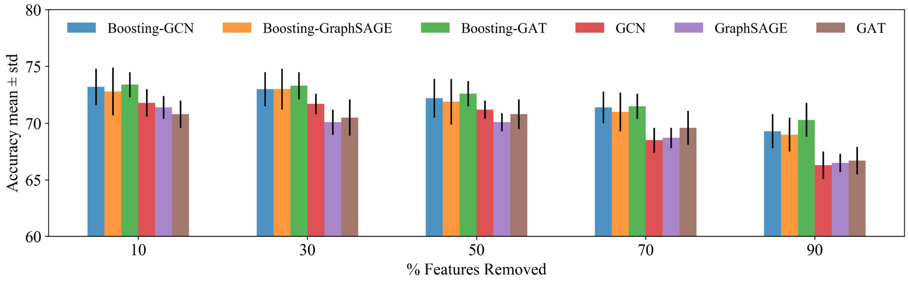


FIGURE 4 | Classification accuracy for the Cora dataset. The features are removed randomly, and the result of 10 runs is averaged. A different random seed is used for every run (i.e., removing different features from each node), but the same 10 random seeds are used across models.

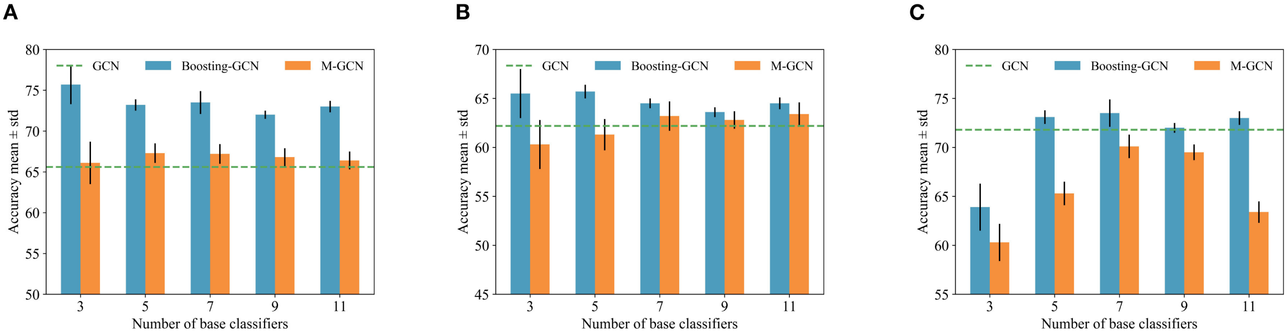


FIGURE 5 | Classification results of Boosting-GCN and M-GCN with different base classifiers. **(A)** Cora, **(B)** Citeseer, and **(C)** Pubmed.

4.7. Why Ensemble Method Useful?

This section analyzes why the ensemble learning approach works on imbalanced datasets and the advantages of Boosting-GNN over traditional GNN. The process of ensemble learning can be divided into two steps:

1) Generating multiple base classifiers for integration. Our model could adjust the weight of samples, adopt specific strategies to reconstruct the dataset, and assign smaller weights to the determined samples and larger weights to the uncertain samples. It makes subsequent base classifiers focus more on samples that are difficult to be classified. In general, the samples of minority classes in imbalanced datasets are more likely to be misclassified. By changing the weights of these samples, subsequent base classifiers can focus more on these samples.
2) Combining the results of the base classifiers. The weight of the classifier is obtained according to the error of the classifier. The base classifier with high classification accuracy has greater weight and a greater influence on the final combined classifier. In contrast, the base classifier with low classification accuracy has less weight and impact on the final combined classifier.

We independently trained M GCNs using the same strategy described in Equation (11) and named this method M-GCN. We compare Boosting-GNN with M-GCN, which is trained according to the hard voting frameworks. Using the synthetic imbalanced datasets in Section 4.3, we changed M and conducted several experiments. Ten runs with different random seeds were conducted to calculate the mean and SD. The experimental results are shown in **Figure 5**, and the classification results of GCN are represented by dotted lines. By effectively setting the number of base classifiers, Boosting-GCN significantly improves classification accuracy compared with M-GCN and GCN.

Next, in order to study the misclassification of samples, we observed the confusion matrix. To increase the imbalance, s is set to 5. The last class is selected as the majority class, and the other classes are selected as the minority classes for convenience. Ten experiments are conducted, and the confusion matrix of the average experimental results is shown in **Figure 6**. Compared with the confusion matrix of the classification performed by GCN, Boosting-GCN achieves a better classification performance.

Due to the sample imbalance, the classifier tends to divide the samples into the majority class, which is reflected by the fact that the last column of the confusion matrix usually has the maximum value (with the brightest color). Compared with GNN, Boosting-GNN improves the performance to a certain extent, especially on the Cora dataset. Based on the aggregation of base estimators, the values on the diagonal of the confusion matrix increase, and the values in the last column of the confusion matrix decrease.

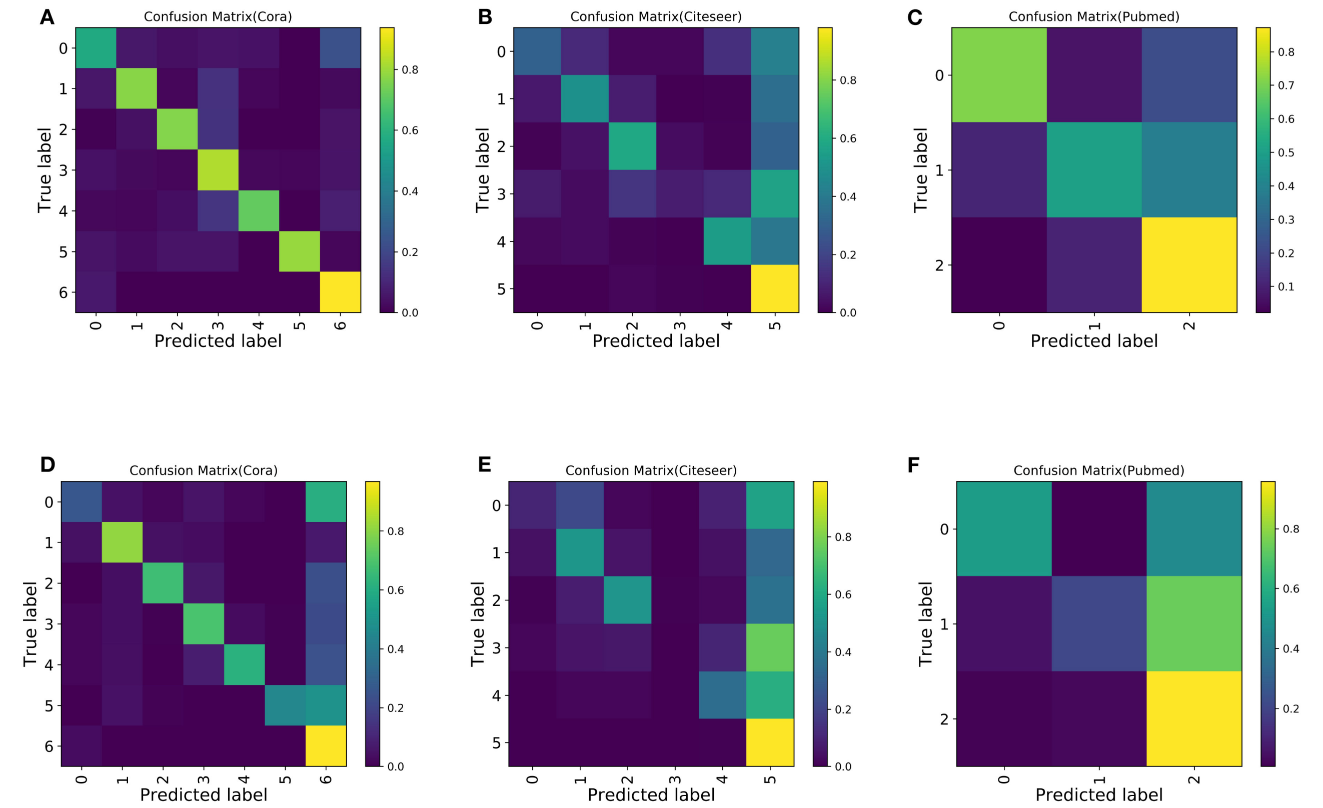


FIGURE 6 | Confusion matrix for the Cora, Pubmed, and Citeseer datasets with 30 labeled nodes for majority class and five labeled nodes for the minority class. **(A)** Boosting-GCN on Cora. **(B)** Boosting-GCN on Citeseer. **(C)** Boosting-GCN on Pubmed. **(D)** GCN on Cora. **(E)** GCN on Citeseer. **(F)** GCN on Pubmed.

TABLE 4 | Comparison of running time when using different number of GCN base classifiers.

Method	5-classifier	7-classifier	9-classifier
M-GCN	28.76 s	39.52 s	51.04 s
Boosting-GCN-t	10.44 s	13.43 s	18.03 s
Boosting-GCN-w/o	18.36 s	27.64 s	34.83 s

We use Cora and train each base classifier for 100 epochs.

In summary, Boosting-GNN integrates multiple GNN classifiers to reduce the effect of overfitting to a certain degree. Moreover, Boosting-GNN reduces the deviation caused by a single classifier and achieves better robustness. Boosting-GNN is an improvement of traditional GNN and makes AdaBoost compatible with GNN. Boosting-GNN achieves higher classification accuracy than a single GNN on imbalanced datasets with the same number of learning epochs.

4.8. Analysis of Training Time

In this section, we conduct a time-consuming analysis of the experiment. We measure the training time on an NVIDIA Tesla V100 GPU. The time required to train the original GCN model for 100 epochs is 6.11s. The time consumed by M-GCN and Boosting-GCN is shown in the **Table 4**. Boosting-GCN-t and Boosting-GCN-w/o denote Boosting-GCN with transfer learning and Boosting-GCN without migration learning, respectively.

Compared to GCN, Boosting-GCN consumes exponentially more time. However, Boosting-GCN reduces the training time by about 50% compared to M-GCN. The application of transfer learning can significantly reduce the time consumed, and models can achieve similar accuracy.

5. CONCLUSION

A multi-class AdaBoost for GNN, called Boosting-GNN, is proposed in this article. In Boosting-GNN, several GNNs are used as base estimators, which are trained sequentially. Also, the errors of a previous GNN are used to update the weights of samples for the next GNN to improve performance. The weights of training samples are incorporated in to the cross-entropy error function in the GNN back propagation learning algorithm. The appliance of transfer learning can significantly reduce the time consumed for computation. The performance of the proposed Boosting-GNN for processing imbalanced data is tested. The experimental results show that Boosting-GNN achieves better performance than state-of-the-arts on synthetic imbalanced datasets, with an average performance improvement of 4.5%.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GitHub, https://github.com/tkipf/gcn/tree/master/gcn/data.

AUTHOR CONTRIBUTIONS

SS performed the data analyses and wrote the manuscript. KQ and SY designed the algorithm. LW and JC analyzed the data. BY did supervision and project administration. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the National Key R&D Program of China under Grant No. 2017YFB1002502 and National Natural Science Foundation of China (Nos. 61701089, 61601518, and 61372172).

REFERENCES

Abu-El-Haija, S., Kapoor, A., Perozzi, B., and Lee, J. (2019). "N-gcn: Multi-scale graph convolution for semi-supervised node classification," in *UAI* (Tel Aviv-Yafo).

Bai, L., Yao, L., Li, C., Wang, X., and Wang, C. (2020). Adaptive graph convolutional recurrent network for traffic forecasting. *ArXiv, abs/2007.02842.*

Blaszczynski, J., and Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing* 150, 529–542. doi: 10.1016/j.neucom.2014.07.064

Buda, M., Maki, A., and Mazurowski, M. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw*. 106, 249–259. doi: 10.1016/j.neunet.2018.07.011

Byrd, J., and Lipton, Z. C. (2019). "What is the effect of importance weighting in deep learning?" in *ICML* (Long Beach, CA).

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. (2010). "Toward an architecture for never-ending language learning," in *AAAI* (Atlanta, GA).

Chawla, N. V., Bowyer, K., Hall, L., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chawla, N. V., Lazarevic, A., Hall, L., and Bowyer, K. (2003). "Smoteboost: Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science, Vol. 2838*, eds N. Lavra, D. Gamberger, L. Todorovski, and H. Blockeel (Berlin; Heidelberg: Springer).

Chen, L., Wu, L., Hong, R., Zhang, K., and Wang, M. (2020). "Revisiting graph based collaborative filtering: a linear residual graph convolutional network approach," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, NY), 27–34.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. J. (2019). "Class-balanced loss based on effective number of samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 9260–9269.

Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. J. (2018). "Large scale fine-grained categorization and domain-specific transfer learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4109–4118.

Douzas, G., and Baç ao, F. (2019). Geometric smote a geometrically enhanced drop-in replacement for smote. *Inf. Sci.* 501, 118–135. doi: 10.1016/j.ins.2019.06.007

Galar, M., Fernández, A., Tartas, E. B., and Herrera, F. (2013). Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognit*. 46, 3460–3471. doi: 10.1016/j.patcog.2013.05.006

Gupta, A., Dollár, P., and Girshick, R. B. (2019). "Lvis: a dataset for large vocabulary instance segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 5351–5359.

Hai-xiang, G., Yi-jing, L., Ya-nan, L., Xiao, L., and Jin-ling, L. (2016). Bpso-adaboost-knn ensemble learning algorithm for multi-class imbalanced data classification. *Eng. Appl. Artif. Intell.* 49, 176–193. doi: 10.1016/j.engappai.2015.09.011

Hamilton, W. L., Ying, Z., and Leskovec, J. (2017). "Inductive representation learning on large graphs," in *NIPS* (Long Beach, CA).

Han, H., Wang, W., and Mao, B. (2005). "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, Vol. 3644*, eds D. S. Huang, X. P. Zhang, and G. B. Huang (Berlin; Heidelberg: Springer).

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Stat. Interface*. 2, 349–360. doi: 10.4310/SII.2009.v2.n3.a8

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "Adasyn: adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong: IEEE), 1322–1328.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778.

Huang, W., Yan, H., Cheng, K., Wang, C., Li, J., Wang, Y., et al. (2021a). A neural decoding algorithm that generates language from visual activity evoked by natural images. *Neural Netw*. 144, 90–100. doi: 10.1016/j.neunet.2021.08.006

Huang, W., Yan, H., Cheng, K., Wang, Y., Wang, C., Li, J., et al. (2021b). A dual–channel language decoding from brain activity with progressive transfer training. *Hum. Brain Mapp*. 42, 5089–5100. doi: 10.1002/hbm.25603

Japkowicz, N., and Stephen, S. (2002). The class imbalance problem: a systematic study. *Intell. Data Anal*. 6, 429–449. doi: 10.3233/IDA-2002-6504

Khoshgoftaar, T., Fazelpour, A., Dittman, D., and Napolitano, A. (2015). "Ensemble vs. data sampling: which option is best suited to improve classification performance of imbalanced bioinformatics data?" in *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)* (Vietri sul Mare, SA), 705–712.

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *ICLR, abs/1609.02907*.

Li, G., Müller, M., Thabet, A. K., and Ghanem, B. (2019). "Deepgcns: can gcns go as deep as cnns?" in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul), 9266–9275.

Li, J., Peng, H., Cao, Y., Dou, Y., Zhang, H., Yu, P. S., et al. (2021). Higher-order attribute-enhancing heterogeneous graph neural networks. *ArXiv, abs/2104.07892*. doi: 10.1109/TKDE.2021.3074654

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE), 2999–3007.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: common objects in context. *ECCV, abs/1405.0312*. doi: 10.1007/978-3-319-10602-1_48

Lv, Q., Feng, W., Quan, Y., Dauphin, G., Gao, L., and dao Xing, M. (2021). Enhanced-random-feature-subspace-based ensemble cnn for the imbalanced hyperspectral image classification. *IEEE J. Select. Top. Appl. Earth Observat. Remote Sens*. 14, 3988–3999. doi: 10.1109/JSTARS.2021.3069013

Nanni, L., Fantozzi, C., and Lazzarini, N. (2015). Coupling different methods for overcoming the class imbalance problem. *Neurocomputing* 158, 48–61. doi: 10.1016/j.neucom.2015.01.068

Peng, H., Zhang, R., Dou, Y., Yang, R., Zhang, J., and Yu, P. S. (2021). Reinforced neighborhood selection guided multi-relational graph neural networks. *ArXiv, abs*/2104.07886.

Ramentol, E., Mota, Y., Bello, R., and Herrera, F. (2011). Smote-rsb*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowl. Inf. Syst*. 33, 245–265. doi: 10.1007/s10115-011-0465-6

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis*. 115, 211–252. doi: 10.1007/s11263-015-0816-y

Seiffert, C., Khoshgoftaar, T., Hulse, J., and Napolitano, A. (2010). Rusboost: a hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. A Syst. Hum*. 40, 185–197. doi: 10.1109/TSMCA.2009.2029559

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. *AI Mag*. 29, 93–106. doi: 10.1609/aimag.v29i3.2157

Shen, L., and Lin, Z. (2016). "Relay backpropagation for effective learning of deep convolutional neural networks," in *Computer Vision–ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, Vol. 9911*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling (Cham: Springer), 467–482.

Sun, K., Lin, Z., and Zhu, Z. (2021). Adagcn: adaboosting graph convolutional networks into deep models. *ArXiv, abs/1908.05081*.

Taherkhani, A., Cosma, G., and Mcginnity, T. (2020). Adaboost-cnn: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing* 404, 351–366. doi: 10.1016/j.neucom.2020.03.064

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. *ICLR, abs*/1710.10903.

Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P., et al. (2019). "Heterogeneous graph attention network," in *The World Wide Web Conference* (San Francisco, CA).

Wang, Y.-X., Ramanan, D., and Hebert, M. (2017). "Learning to model the tail," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (Long Beach, CA).

Wei, X.-S., Cui, Q., Yang, L., Wang, P., and Liu, L. (2019). Rpc: a large-scale retail product checkout dataset. *ArXiv, abs/1901.07249*.

Wu, F., Zhang, T., de Souza, A. H., Fifty, C., Yu, T., and Weinberger, K. Q. (2019). Simplifying graph convolutional networks. *ArXiv, abs/1902.07153*.

Yang, Z., Cohen, W. W., and Salakhutdinov, R. (2016). Revisiting semi-supervised learning with graph embeddings. *ICML, abs/1603.08861*.

Yin, X., Yu, X., Sohn, K., Liu, X., and Chandraker, M. (2019). "Feature transfer learning for face recognition with under-represented data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA), 5697–5706.

Yu, D., Xiong, W., Droppo, J., Stolcke, A., Ye, G., Li, J., et al. (2016). Deep convolutional neural networks with layer-wise context expansion and attention. *Proc. Interspeech* 2016, 17–21. doi: 10.21437/Interspeech.2016-251

Yu, W., and Qin, Z. (2020). Graph convolutional network for recommendation with low-pass collaborative filters. *ArXiv, abs/2006.15516*.

Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell*. 40, 1452–1464. doi: 10.1109/TPAMI.2017.2723009

Zhou, Q., Sun, B., Song, Y., and Li, S. (2020). "K-means clustering based undersampling for lower back pain data," in *Proceedings of the 2020 3rd International Conference on Big Data Technologies* (Qingdao).

Zitnik, M., and Leskovec, J. (2017). Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* 33, i190–i198. doi: 10.1093/bioinformatics/btx252

Zou, Y., Yu, Z., Kumar, B. V., and Wang, J. (2018). "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *ECCV* (Munich).

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 09 December 2021
doi: 10.3389/fnbot.2021.784053

Improving the Transferability of Adversarial Examples With a Noise Data Enhancement Framework and Random Erasing

*Pengfei Xie, Shuhao Shi, Shuai Yang, Kai Qiao, Ningning Liang, Linyuan Wang, Jian Chen, Guoen Hu and Bin Yan**

Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategy Support Force Information Engineering University, Zhengzhou, China

OPEN ACCESS

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Qilong Zhang,
University of Electronic Science and Technology of China, China
Xiaosen Wang,
Huazhong University of Science and Technology, China

****Correspondence:***
Bin Yan
ybspace@hotmail.com

Received: *27 September 2021*
Accepted: *04 November 2021*
Published: *09 December 2021*

Citation:
Xie P, Shi S, Yang S, Qiao K, Liang N, Wang L, Chen J, Hu G and Yan B (2021) Improving the Transferability of Adversarial Examples With a Noise Data Enhancement Framework and Random Erasing. Front. Neurorobot. 15:784053. doi: 10.3389/fnbot.2021.784053

Deep neural networks (DNNs) are proven vulnerable to attack against adversarial examples. Black-box transfer attacks pose a massive threat to AI applications without accessing target models. At present, the most effective black-box attack methods mainly adopt data enhancement methods, such as input transformation. Previous data enhancement frameworks only work on input transformations that satisfy accuracy or loss invariance. However, it does not work for other transformations that do not meet the above conditions, such as the transformation which will lose information. To solve this problem, we propose a new noise data enhancement framework (NDEF), which only transforms adversarial perturbation to avoid the above issues effectively. In addition, we introduce random erasing under this framework to prevent the over-fitting of adversarial examples. Experimental results show that the black-box attack success rate of our method Random Erasing Iterative Fast Gradient Sign Method (REI-FGSM) is 4.2% higher than DI-FGSM in six models on average and 6.6% higher than DI-FGSM in three defense models. REI-FGSM can combine with other methods to achieve excellent performance. The attack performance of SI-FGSM can be improved by 22.9% on average when combined with REI-FGSM. Besides, our combined version with DI-TI-MI-FGSM, i.e., DI-TI-MI-REI-FGSM can achieve an average attack success rate of 97.0% against three ensemble adversarial training models, which is greater than the current gradient iterative attack method. We also introduce Gaussian blur to prove the compatibility of our framework.

Keywords: adversarial examples, black-box attack, transfer-based attack, data enhancement, transferability

1. INTRODUCTION

In recent years, the data-driven deep neural network (DNNs) has developed rapidly due to its excellent performance. It has made outstanding achievements in image classification (He et al., 2016; Szegedy et al., 2017), target detection (Redmon and Farhadi, 2018; Bochkovskiy et al., 2020), face recognition (Deng et al., 2019), automatic driving (Bojarski et al., 2016), natural language processing (Gehring et al., 2017; Vaswani et al., 2017) and so on. Unfortunately, the current deep learning model has been proved to be not robust, and they are vulnerable to adversarial examples. In the field of computer vision, adversarial examples are specially tailored to the target model, which can make the model misclassified but are visually similar to the original sample. Subsequently, with the development of adversarial attack and defense, its attack range is gradually expanded to speech recognition model (Carlini and Wagner, 2018), reinforcement learning model (Behzadan and Munir, 2017), graph neural network (Dai et al., 2018), etc.

The adversarial attack was first proposed by Szeged (Szegedy et al., 2013), and they use the L-BFGS optimization algorithm to find adversarial examples. Later, DeepFool (Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017) and other optimization-based algorithms are proposed, but they focus on meeting established optimization goals in white-box attacks. However, these optimization-based methods take too much time and have poor transferability in black-box attacks. A black-box attack refers to the attack that attacker cannot know the network structure, parameters, and other information of the attacked model. Black-box attacks can be divided into three categories: scores-based, decision-based, and transfer-based attacks. In this paper, we discuss the more difficult black-box transfer attacks. Papernot et al. (2016) find that adversarial examples generated by one model can attack another model. The transferability of adversarial examples is similar to the generalization of model training. The latter is to train a robust model to classify the samples correctly, and the former is to train a robust sample so that it can successfully attack various models. Tramér et al. (2017) show that using the integrated model can train robust adversarial examples with stronger attack performance. However, simply adding pre-models requires a lot of storage space and time cost; hence researchers turn their attention to data enhancement, such as Dong et al. (2019), Lin et al. (2019), and Xie et al. (2019). These works essentially make use of the translation invariance, resize invariance, scaling invariance, and other properties of convolutional neural network (CNN), but when it exceeds a certain transformation range, the above properties will not hold, and the method based on data enhancement will fail. Based on this problem, we propose a NDEF, which solves the problem of limited change range. Specifically, we only perform input transformations against adversarial perturbations instead of the entire image. This avoids the trouble of misclassification of the original image in a wide range of changes. In addition, inspired from Zhong et al. (2020), we introduce a new data enhancement method in this framework, namely random erasing, which can effectively avoid the adversarial examples falling into an over-fitting state. Experiments show that the average success rate of our method is 4.2% higher than DI-FGSM and 2.5% higher than SI-FGSM on average, and DI-TI-MI-FGSM combined with our method can achieve an average attack success rate of 97.0% against three ensemble adversarial training models.

Our main contributions are summarized as follows.

- We propose a noise data enhancement framework (NDEF), which effectively solves the problem that some transformations, such as random erasing and Gaussian blur, that do not satisfy accuracy invariance cannot work in the previous framework. These input transformation methods can work in our framework.
- We introduce random erasing as an input transform into the gradient iterative attack for the first time and call it Random Erasing Iterative Fast Gradient Sign Method (REI-FGSM). The experimental results show that the attack success rate of our method is 4.2% higher than DI-FGSM and 2.5% higher than SI-FGSM on average. Our method can be combined with other gradient iteration methods. DI-TI-MI-REI-FGSM can achieve an average attack success rate of 97.0% against three ensemble adversarial training models, which is greater than the current gradient iterative attack method.

2. RELATED WORK

2.1. Adversarial Attack

Szegedy et al. first produce adversarial examples using box constraint algorithm L-BFGS. However, this method requires huge costs; hence (Goodfellow et al., 2015) propose a FGSM to generate adversarial examples. This method belongs to the one-step iterative attack method, aiming to find the direction of maximizing the loss function. Subsequently, Kurakin et al. (2016) propose a multistep iterative attack method I-FGSM based on FGSM, which can ensure that the adversarial examples can find the direction of the maximum loss function in each iteration. I-FGSM can achieve excellent performance in white box attack, but the attack performance of black-box is poor. This is because I-FGSM is easy to fall into over-fitting on the substitute model. Therefore, many works begin to study how to improve the transferability of adversarial examples. At present, black-box transfer attacks can be divided into four categories, i.e., based on gradient information mining, based on data enhancement, based on model enhancement, and intermediate-layers attack.

2.1.1. Gradient Information Mining Methods

Gradient information mining methods refer to various methods that attackers deal with gradient after gradient back-iteration to adjust the current gradient, propagation. Dong et al. (2018) propose MI-FGSM, which uses the momentum in the gradient iteration process to stabilize the gradient direction and escape from the local extremum. Similar to MI-FGSM, NI-FGSM (Lin et al., 2019) escapes local extremum faster by introducing Nesterov acceleration gradient. Wang and He (2021) propose variance tuning MI-FGSM, as VMI-FGSM, which uses the gradient variance of the previous iteration to adjust the current gradient, stabilize the update direction, and avoid poor local optimization in the iteration process. Wu et al. (2018) use Gaussian noise to simulate local fluctuations in substitute models to improve transferability. Gao et al. (2020) find that increasing the step size can increase the transferability, but it can lead to gradient overflow; hence, they propose PI-FGSM, which uses pre-trained convolution kernels to project the proposed overflow information to the surrounding area to improve transferability. Wu et al. (2020a) use the skip structure of the residual network to improve the transferability. Specifically, the gradient of the residual network is decomposed, and the attenuation parameter is introduced to reduce the gradient from the residual block and pay more attention to the gradient information flow from the bottom.

2.1.2. Data Enhancement Methods

Data enhancement methods are methods that an attacker performs a series of transformations on a sample before entering a model to enhance transferability. DI-FGSM (Xie et al., 2019) improves the transferability of adversarial examples by introducing random resizing and random padding for input in the gradient iteration process. Using the scale invariance of CNN, SI-FGSM (Lin et al., 2019) introduces scale transformation in the gradient iteration process to improve the transferability of adversarial examples. TI-FGSM (Dong et al., 2019) uses the translation invariance of CNN and replaces the translation operation with pre-trained convolution to save substantial time and space costs. Zou et al. (2020) find that TI-FGSM can be regarded as a Gaussian blur, and the information of normal image will be lost by the Gaussian blur, while the vertical and horizontal stripes can alleviate this phenomenon. They further find that the larger the scaling ratio of DI-FGSM will generate more stripes, which will make the mitigation effect better. Based on this, they propose resized-diverse-inputs methods, which can effectively improve transferability. Wu et al. (2021) train an adversarial transformation network to replace previous transformation algorithms. Specifically, they first train an adversarial transformation network using the maximum and minimum, which can effectively correct the adversarial examples while keeping the original samples unchanged. Then they combine adversarial transformation networks with the target model and attack them. The previous work is to perturb a single image. Wang et al. (2021a) propose Admix Attack Method (AAM), which integrates some information of other categories of images into the original category to enhance transferability.

2.1.3. Model Enhancement Methods

Model enhancement methods refer to the methods by which an attacker improves transferability by model integration or transformation. Liu et al. (2017) propose a model-ensemble attack method that can effectively attack robust black-box models for adversarial training. Li et al. (2020) erode the dropout layer and skip the connection layer of the model to obtain rich network models at low cost and then improve transferability through vertical integration.

2.1.4. Intermediate-Layers Attack Methods

Intermediate-layers attack methods launch attacks by using information from the network middle layer instead of the logit layer. Inkawhich et al. (2020) use the Euclidean distance to reduce the discrepancy between the intermediate source and target features to achieve target attacks, but this pixel-wise Euclidean distance would impose a spatial-consistency constraint

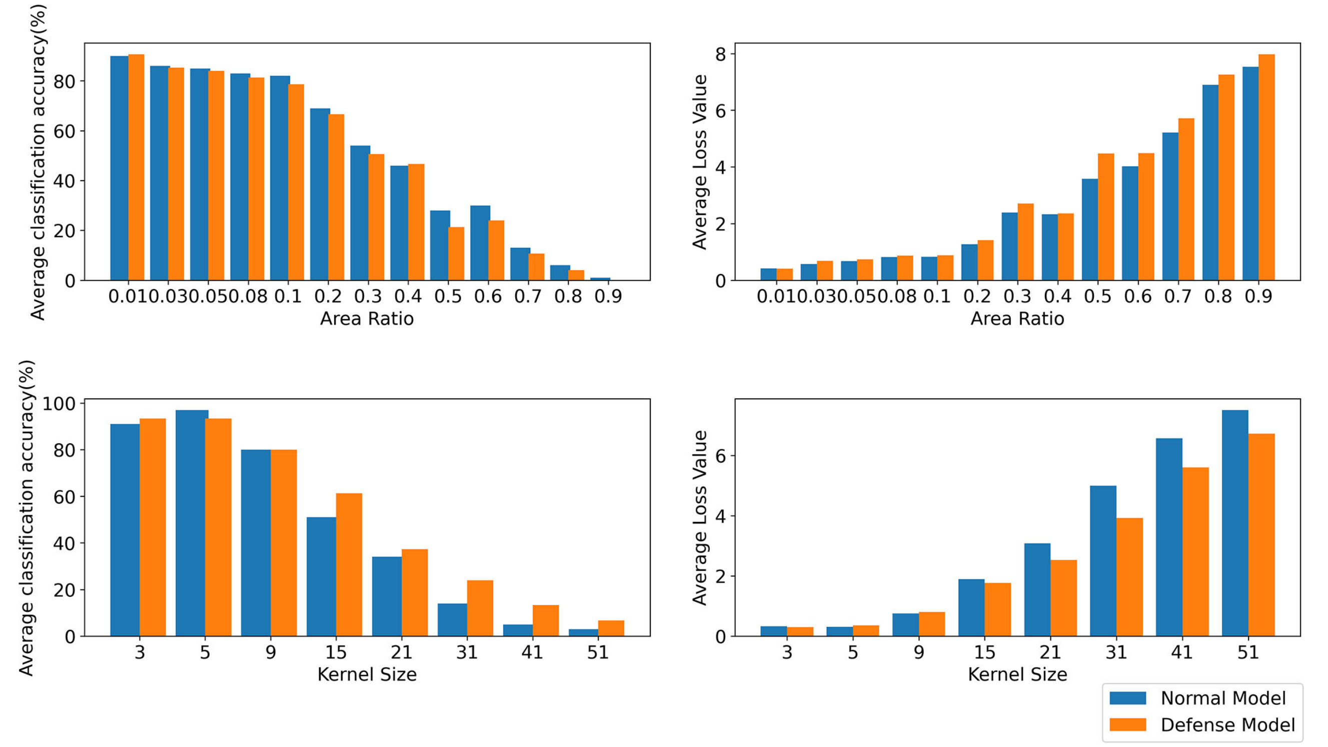


FIGURE 1 | The first line shows the average classification accuracy (%) and average loss value under normal model and defense model with different area ratios by random erasing. The second line shows the average classification accuracy (%) and average loss value under normal model and defense model with different kernel sizes by Gaussian blur. The results are averaged over 1,000 images.

on them. To solve this problem, Gao et al. (2021) propose Pair-wise Alignment Attack (PAA) and Global-wise Alignment Attack (GAA), which use statistic alignment. Specifically, PAA uses maximum mean discrepancy (MMD) to estimate the difference between the intermediate source and target features, while GAA uses mean and variance to achieve this goal. Inkawhich et al. (2020) propose Feature Distribution Attack (FDA), which first trains a binary network to extract the feature distribution of classes and layers. Then they maximize the probability of specific classes in the auxiliary network to accomplish target attack. Wu et al. (2020b) find that the attention regions of different models are almost the same. Based on this, they propose an Attention-guided Transfer Attack (ATA) method, and add the attention region loss into the loss function to make the attention region change more to enhance transferability. Wang et al. (2021b) propose Feature Importance-aware Attack (FIA), which uses a random transformation to destroy the key features that determine the decisions of different models, and then gradient aggregation is carried out to improve transferability.

2.2. Adversarial Defense

Adversarial training is currently considered to be the strongest method defending adversarial examples, which add adversarial examples during model training. These works (Szegedy et al., 2013; Goodfellow et al., 2015) first mention adversarial training. Subsequently, Madry et al. (2019) analyze adversarial training from the perspective of robust optimization for the first time, propose a min-max framework, and use the adversarial examples generated by Project Gradient Descent (PGD) to achieve the approximate solution of the framework. Input transformation is another common defense method. Madry et al. (2019) find that JPEG compression can effectively suppress small perturbation adversarial examples. Xie et al. (2017) mitigates the impact of attacks by random resizing and random padding. In recent years, some works (Raghunathan et al., 2018; Fischer et al., 2020) has begun to focus on certified defense methods.

3. METHODS

3.1. Problem Definition

3.1.1. Adversarial Example

Suppose x is a clean sample, y^{true} is the corresponding real label. For a trained DNN F_1, it can correctly classify samples x as labels y^{true}. By adding a small perturbation δ to the original sample, the adversarial examples $x + \delta$ can make the DNN F_1 misclassified. The generation of the small perturbation is generally obtained by maximizing the loss function $J(x, y^{true}, \theta)$, where θ represents the network structure parameters, and the loss function generally selects the cross entropy loss function.

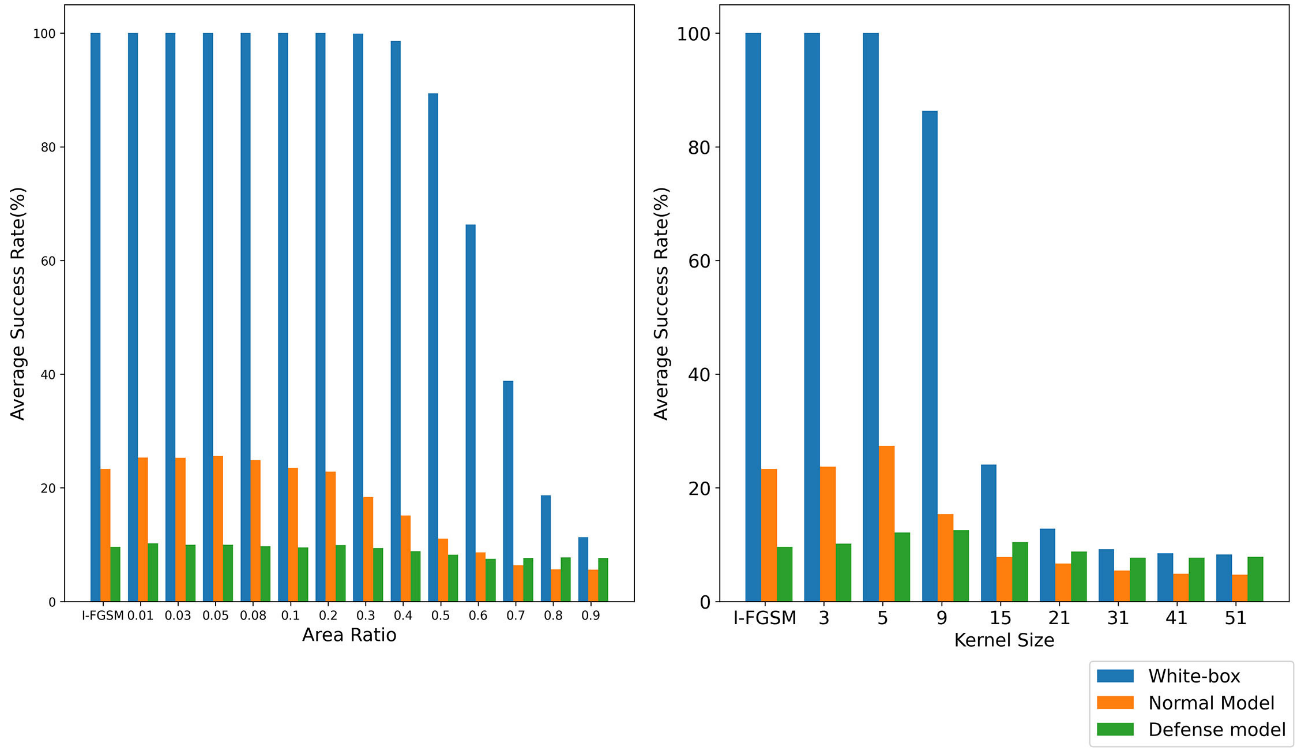


FIGURE 2 | The attack success rate (%) on seven models, the adversarial examples are crafted by REI-FGSM on Inc-v3 model with different area ratios. The attack success rate (%) on seven models, the adversarial examples are crafted by random erasing and Gaussian blur on Inc-v3 model with different area ratios and kernel size in the original framework.

3.1.2. Black-Box Transfer Attack

Assuming DNNs F_1 and F_2 perform the same task, which both can correctly classify clean samples x as labels y^{true}, we denote θ_1 θ_2 are the network parameters of F_1 and F_2 respectively. In the black-box attack background, only the parameters F_1 are known, and the parameters F_2 are unknown. The goal of black-box attack is that the adversarial examples generated by the existing network structure information θ_1 can make misclassification on F_2, i.e., $F_2(x^{adv}) \neq y^{true}$.

3.2. Classical Attack Methods

In this section, we will briefly review the classic adversarial attack algorithms.

Fast Gradient Sign Method: Goodfellow et al. (2015) believe that the linear nature of the neural network leads to the generation of adversarial examples, and propose an FGSM for the first time. The purpose of this method is to find the direction of the maximum loss function. The formula is as follows :

$$x^{adv} = x + \varepsilon \cdot sign(\nabla_x L(x, y^{true}, \theta)) \tag{1}$$

Iterative FGSM (I-FGSM): Kurakin et al. (2016) propose an iterative version of FGSM, i.e., I-FGSM. Compared with FGSM, I-FGSM can more accurately maximize the loss function. The formula is as follows:

$$x_0^{adv} = x \tag{2}$$

$$x_{t+1}^{adv} = Clip_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(\nabla_x L(x_t^{adv}, y^{true}, \theta))\} \tag{3}$$

where α represents the gradient iteration step size, and $Clip_x^{\varepsilon}$ means that the adversarial examples x^{adv} is limited to the norm ball l_∞ of the original sample.

Momentum I-FGSM (MI-FGSM): Dong et al. (2018) introduce momentum into the gradient iteration process to stabilize the gradient update direction and escape from the local extremum. The formula is as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y^{true})}{\left\| \nabla_x J(x_t^{adv}, y^{true}) \right\|_1} \tag{4}$$

$$x_{t+1}^{adv} = Clip_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(g_{t+1})\} \tag{5}$$

where μ represents the attenuation factor.

Diverse Input Iterative FGSM (DI-FGSM): Xie et al. (2019) improve the transferability of adversarial examples by introducing input transformation. The method is as follows:

$$x_{t+1}^{adv} = Clip_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(\nabla_{x_t^{adv}} J(D(x_t^{adv}, p), y^{true}))\} \tag{6}$$

where D represents the input transformation, and p represents the transformation probability.

Translation-Invariant Attack Method (TI-FGSM): Dong et al. use the translation invariance of CNN and replace translation operations with convolution kernels to improve the transferability of adversarial examples.

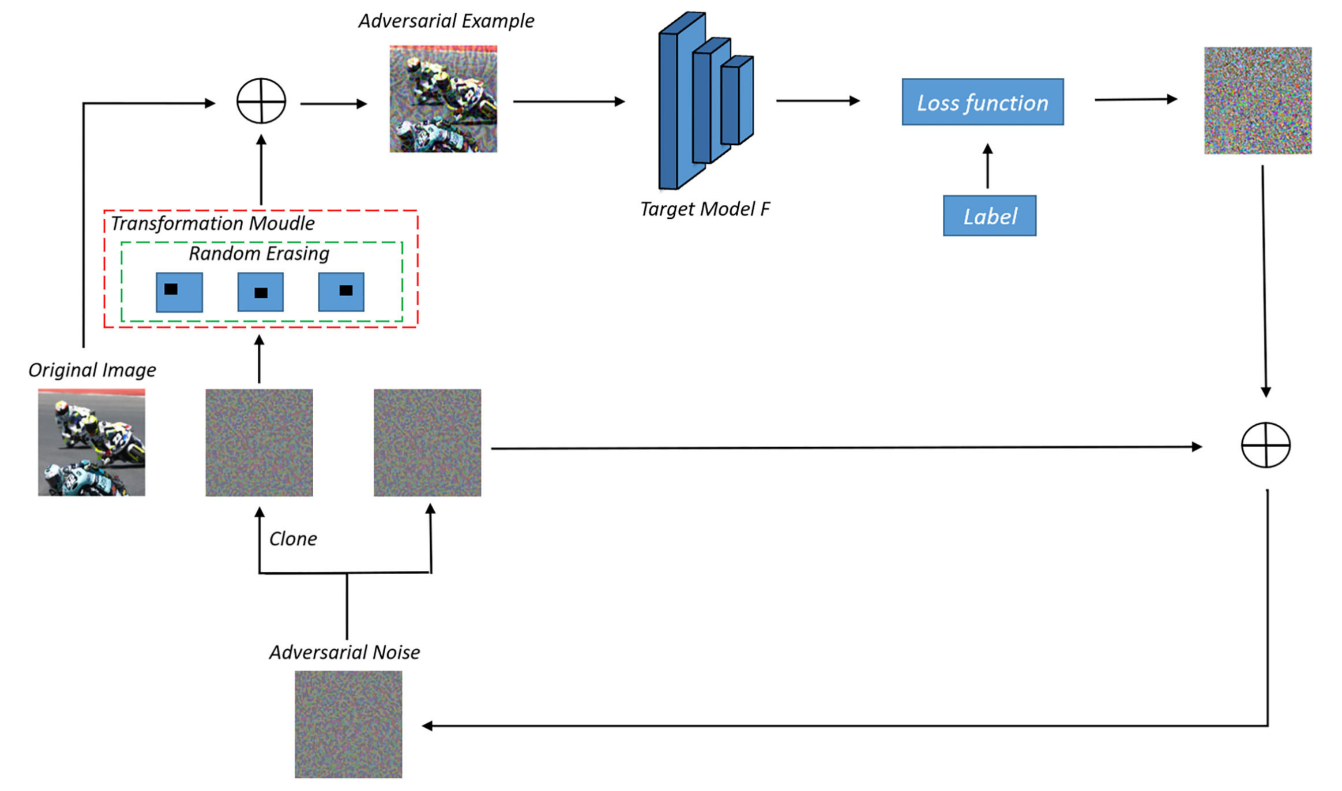


FIGURE 3 | The framework of our methods.

3.3. Motivation

It is difficult to obtain good transferability by simply maximizing the loss function, such as the classical algorithm I-FGSM, because the adversarial examples generated by these methods are very easy to fall into overfitting on the substitute model in the gradient iteration process. Studies (Dong et al., 2019; Lin et al., 2019; Xie et al., 2019) have shown that the input transformation of the whole image can increase the transferability of adversarial examples. The precondition of this method is that the input transformation must satisfy certain precision invariance or loss invariance (Lin et al., 2019; Liu and Li, 2020). However, for some data enhancement methods that may lose some information, too large a transformation scale makes them unable to adapt to the above framework. We give an intuitive example by random erasing and Gaussian blur. Specifically, for random erasing, we randomly generate matrices with different area ratios from 0.01,0.03,0.05,0.08,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8, and 0.9 and set the pixel value in the matrix to 0. For Gaussian blur, we use different kernel sizes from 3,5,9,15,21,31,41, and 51 to blur the original sample. As shown in **Figure 1**, the first line is the classification accuracy and loss value after random erasing, and the second line is the classification accuracy and

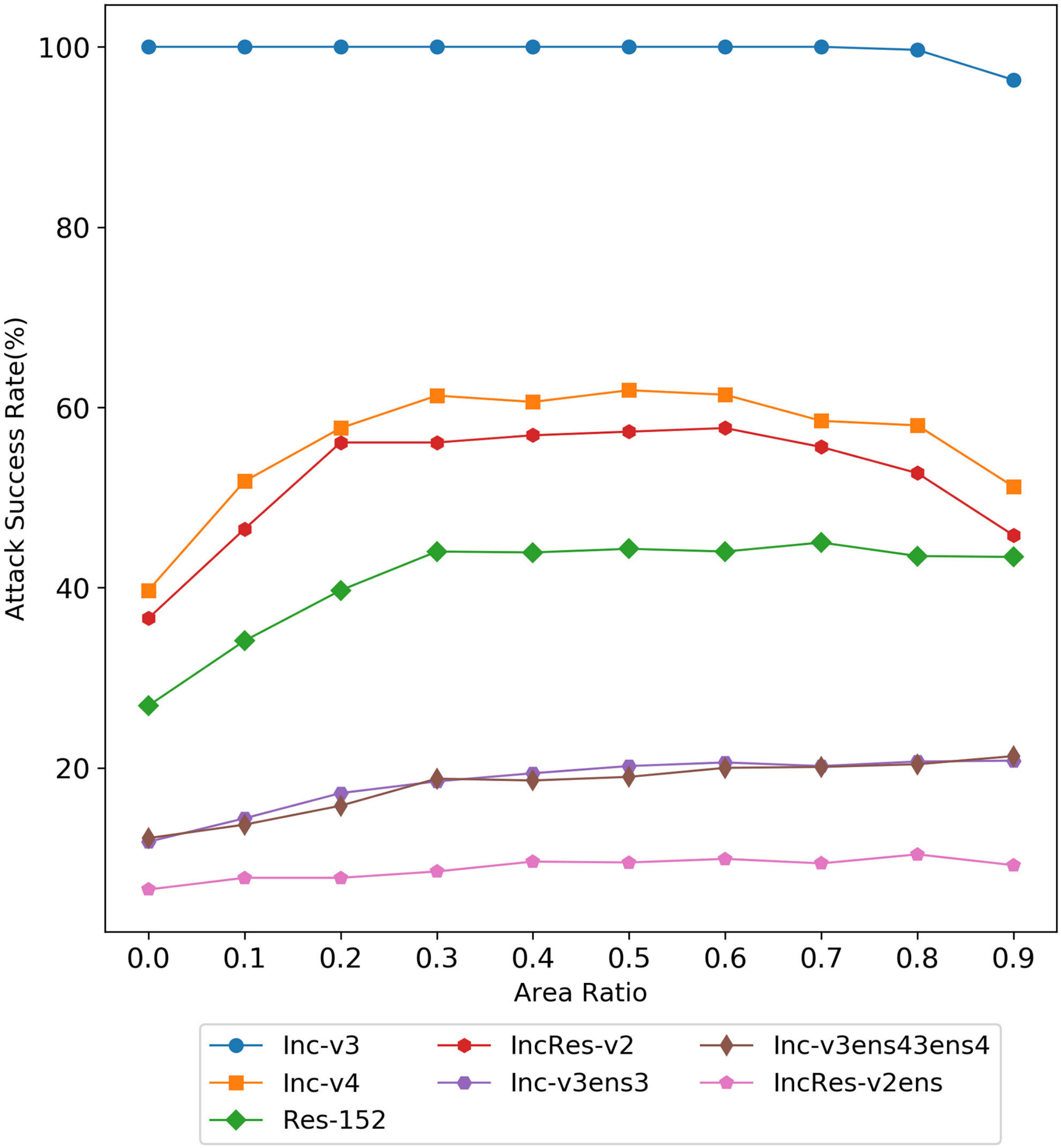


FIGURE 4 | The attack success rate (%) on seven models, the adversarial examples are crafted by Random Erasing Iterative Fast Gradient Sign Method (REI-FGSM) on Inc-v3 model with different area ratios.

loss value after Gaussian blur. It can be seen that when the area ratio is greater than 0.2 and the kernel size is greater than 9, the classification accuracy of CNN decreases significantly. Then, in the original framework, we test the attack success rate of random erasing and Gaussian blur under different transformation scales. As shown in **Figure 2**, the experimental results show that when the area ratio is greater than 0.05, the black-box attack success rate decreases. When the area rate is greater than 0.4, the black-box attack success rate decreases significantly. For Gaussian blur, when the Gaussian kernel is greater than 9, the black box attack rate decreases cliff-like. The experimental results show that the previous framework does not apply to some data enhancement methods with too large transformation scale. Based on this problem, we propose a noise data enhancement framework. Since our framework only transforms against perturbation, the structure information of the original sample will not be destroyed, which can maintain the accuracy invariance. In addition, the transformation of adversarial perturbation can hinder the generation of adversarial examples and prevent over-fitting. Our framework is a supplement to the previous framework, which can mine the potential of some transformation methods without accuracy invariance in transfer attack methods. In this paper, we mainly introduce random erasing. As far as we know, it is the first time that random erasing has been introduced into a transfer attack as an input transformation. Random erasing is an effective data enhancement method. Specifically, the rectangular region of the image is randomly selected, and the pixels are erased or replaced by other values. The generation of adversarial examples with occlusion levels will reduce the risk of overfitting and make the adversarial examples robust to occlusion. In addition, in order to verify that our framework can also be compatible with other methods, we briefly introduce Gaussian blur.

TABLE 1 | The attack success rate (%) of seven models, the leftmost column represents the number of erased matrices whose erased area ratio is 0.1, adversarial examples crafted by REI-FGSM on Inc-v3 model ("*" indicate the white box attack).

Area_number	Inc-v3	Inc-v4	Res-152	IncRes-v2	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens
1	**100.0***	51.8	34.1	46.5	14.4	13.7	7.8
3	**100.0***	62.2	43.7	56.3	17.2	17.2	9.0
5	**100.0***	67.2	48.9	60.7	22.2	17.5	9.6
8	**100.0***	**69.2**	51.7	65.4	23.0	21.1	10.8
10	**100.0***	67.9	**52.3**	**65.7**	22.5	21.5	10.3
15	**100.0***	66.4	50.3	62.7	**23.9**	**22.5**	**10.9**
20	99.9*	64.5	48.6	59.8	21.9	21.9	10.8

The bold value represents the highest success rate for different attack methods under the same experimental conditions.

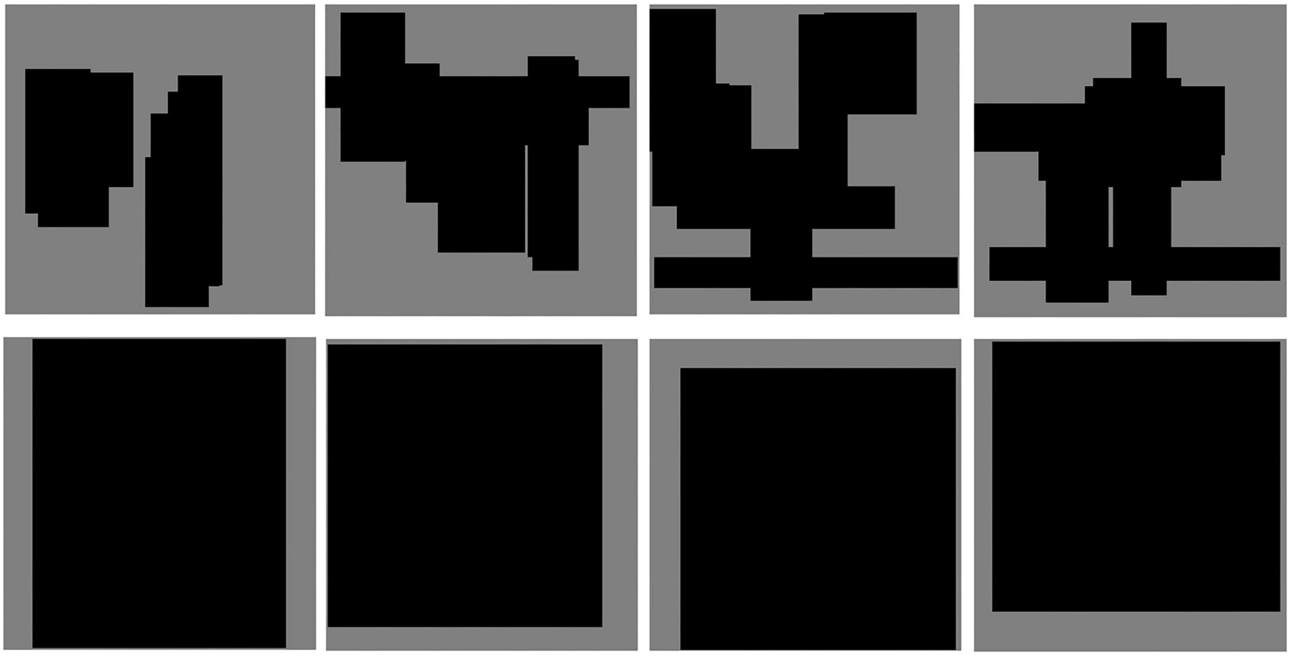


FIGURE 5 | Comparison of multi-matrix erasing (top) and single-matrix erasing (bottom).

Algorithm 1: REI-FGSM

```
Input : An original image $x$, normalized to $[-1, 1]$ and
        corresponding true labels $y^{true}$; maximum
        perturbation value $\varepsilon$; iteration rounds $T$;
        adversarial perturbation $\delta_t$, input image size $W$,
        $H$; lower bound $\theta_L$, upper bound $\theta_H$ of mask
        matrix area ratio; number of matrices $K$.
Output: An adversarial example $x_{adv}$.
$a = \frac{\varepsilon}{T}$;
Initialize $x_0^{adv} = x$;
Random initialization adversarial perturbation $\delta_0$;
for $t \leftarrow 0$ to $T - 1$ do
    Replicate adversarial perturbation $\delta_t$ and get
    adversarial perturbation$\delta_t$*;
    Get the area ratio of random masking matrix
    $\theta_e = Rand(\theta_L, \theta_H)$;
    Get the area of random masking matrix
    $S_e = W * H * \theta_e$;
    for $i \leftarrow 0$ to $K - 1$ do
        if $random(1) > 0.5$ then
            Get the aspect ratio of the $jth$ matrix
        $\varphi_e = Rand(\theta_e, 1)$;
        else:
            Get the aspect ratio of the $jth$ matrix
        $\varphi_e = Rand(1, \frac{1}{\theta_e})$;
        Get the $jth$ matrix length $H_j = Floor(\sqrt{\frac{S_e}{\varphi_j}})$;
        Get the $jth$ matrix width $W_j = Floor(\sqrt{S_e * \varphi_j})$;
        Get the horizontal ordinate of starting pixels of $jth$
        matrix $X_j = Rand(0, (H - H_j))$;
        Get the ordinate of starting pixels of $jth$ matrix
        $Y_j = Rand(0, (W - W_j))$;
        Set 0 for region $[X_j + H_j, Y_j + W_j]$ in $\delta_t$*;
    end
    Calculate gradient $\nabla_{\delta_t} J((x + \delta_t^*), y^{true})$;
    Update adversarial perturbation
    $\delta_t = \delta_t + \alpha \cdot \text{sign}(\nabla_{\delta_t} J((x + \delta_t^*), y^{true}))$;
    Clip the adversarial perturbation $\delta_t = Clip(\delta_t, -\varepsilon, \varepsilon)$;
    Get adversarial examples $x_t^{adv} = x + \delta_t$;
    Clip the adversarial examples
    $x_t^{adv} = Clip(x_t^{adv}, -1, 1)$;
    Get adversarial perturbation $\delta_t = x_t^{adv} - x$;
end
Return $x_t^{adv} = x + \delta_t$;
```

Algorithm 2: GBI-FGSM

```
Input : An original image $x$, normalized to $[-1, 1]$ and
        corresponding true labels $y^{true}$; maximum
        perturbation value $\varepsilon$; iteration rounds $T$;
        adversarial perturbation $\delta_t$; the kernel size $k$;
        Output: An adversarial example $x_{adv}$.
Output: An adversarial example $x_{adv}$.
$a = \frac{\varepsilon}{T}$;
Initialize $x_0^{adv} = x$;
Random initialization adversarial perturbation $\delta_0$;
for $t \leftarrow 0$ to $T - 1$ do
    Replicate adversarial perturbation $\delta_t$ and get
    adversarial perturbation $\delta_t$*;
    Gaussian blur for adversarial perturbation and
    update $\delta_t^* = Gaussianblur\left(\delta_t^*, k\right)$;
    Calculate gradient $\nabla_{\delta_t} J((x + \delta_t^*), y^{true})$;
    Update adversarial perturbation
    $\delta_t = \delta_t + \alpha \cdot \text{sign}(\nabla_{\delta_t} J((x + \delta_t^*), y^{true}))$;
    Clip the adversarial perturbation $\delta_t = Clip(\delta_t, -\varepsilon, \varepsilon)$;
    Get adversarial examples $x_t^{adv} = x + \delta_t$;
    Clip the adversarial examples
    $x_t^{adv} = Clip(x_t^{adv}, -1, 1)$;
    Get adversarial perturbation $\delta_t = x_t^{adv} - x$;
end
Return $x_t^{adv} = x + \delta_t$;
```

3.4. Framework

As far as we know, the current data-enhanced attack methods generally have to satisfy the invariance property as follows:

$$\arg\max((F_{Logit}(x)) = \arg\max(F_{Logit}(T(x))) \tag{7}$$

Meanwhile, input transformation destroys the structure of the adversarial example to remove or weaken its attack performance, which can effectively enhance the diversity of model output. This can be described as the following formula:

$$F_{Logit}(x^{adv}) \neq F_{Logit}(T(x^{adv})) \tag{8}$$

where $T(\cdot)$ represents a certain transformation and F_{Logit} represents the logit output of the model. Lin et al. (2019) and Liu and Li (2020) interpret that model augmentation can be achieved by loss-preserving transformation and accuracy-maintained transformation. However, some transformations that do not meet the CNN invariant characteristics will fail in this framework. In order to make these transformations also play their performance, in this paper, we propose a new data enhancement framework, only aimed at adversarial perturbation, and we replace $F_{Logit}(T(x + \delta))$ with $F_{Logit}(x + T(\delta))$, so that the original sample will not be disturbed.

Meanwhile, the input transformation will affect the adversarial perturbation, thus affecting the logit output of the model. The formula is shown below.

$$F_{Logit}(x + T(\delta)) \neq F_{Logit}(x + \delta) \tag{9}$$

We use **M** to represent the model space for the same task; F is a model in this space. Since the adversarial perturbation is interfered by the input transformation, the logit output of F changes. We can find another model F^* in this space to make its logit output approximate to the logit output of F. The formula is shown below.

$$F^*_{Logit}(x + \delta) \approx F_{Logit}(x + T(\delta)) \tag{10}$$

TABLE 2 | The success rate(%) of non-targeted attacks of seven models.

Model	Attacks	Inc-v3	Inc-v4	Res-152	IncRes-v2	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens
Inc-v3	I-FGSM	**100.0***	29.6	19.4	20.3	11.7	12.1	5.5
	DI-FGSM	99.8*	54.2	32.1	43.6	15.0	16.2	7.1
	SI-FGSM	**100.0***	50.5	38.0	44.9	21.6	**21.7**	10.0
	REI-FGSM	99.7*	**56.5**	**39.6**	**48.8**	**23.8**	21.4	**11.3**
Inc-v4	I-FGSM	43.3	**100.0***	25.5	25.3	11.8	13.0	6.6
	DI-FGSM	66.6	**100.0***	39.8	50.4	14.7	17.7	8.4
	SI-FGSM	69.9	**100.0***	**48.1**	55.3	**26.9**	**26.5**	**14.9**
	REI-FGSM	**72.1**	99.8*	46.7	**56.2**	23.8	23.5	14.0
Res-152	I-FGSM	30.7	24.7	99.5*	16.9	13.0	13.3	6.7
	DI-FGSM	**60.0**	**56.5**	99.2*	**49.3**	21.6	21.1	12.9
	SI-FGSM	43.0	36.3	**99.7***	30.6	20.5	19.2	11.6
	REI-FGSM	49.7	45.2	99.0*	40.1	**25.9**	**25.0**	**16.3**
IncRes-v2	I-FGSM	48.2	38.3	25.5	**100.0***	13.7	13.3	8.2
	DI-FGSM	70.2	66.1	47.9	99.2*	19.3	20.2	12.7
	SI-FGSM	71.5	58.4	49.8	**100.0***	**30.6**	**28.8**	**22.5**
	REI-FGSM	**72.9**	**66.8**	**51.1**	99.2*	30.3	28.3	**22.5**

The top row models are substitute models, and we use them to generate adversarial examples by I-FGSM, DI-FGSM, SI-FGSM, and REI-FGSM ("" indicates the white-box attack). The bold value represents the highest success rate for different attack methods under the same experimental conditions.*

In other words, we use the above framework to change the logit output of the substitute model during each iteration to achieve model augmentation. Our frame diagram is shown in **Figure 3**. Specifically, we copy the adversarial perturbation, one for storing the previous adversarial perturbation information, and one for data enhancement. Here, we introduce random erasing. We study single matrix erasing and multi-matrix erasing, respectively. Specifically, we select randomly the area ratio within a finite interval in each iteration, then select randomly the aspect ratio within the interval confirmed by the area ratio, finally, initialize the starting point of the matrix randomly. The pixels of the matrix can be set to 0, or other values. In this paper, we set the pixel of the erased matrix to 0. The specific algorithm is shown in **Algorithm 1**. In addition, our framework can also be combined with previous methods for the whole image enhancement.

To further verify that our framework can be combined with other algorithms, we introduce Gaussian blur (Gedraite and Hadad, 2011) and call it the Gaussian Blur Iterative FGSM (GBI-FGSM). We prove that using Gaussian blur on the previous framework is not very good, while Gaussian blur in our framework can get relatively good performance, especially on defense models. This is because Gaussian blur in the original framework will lose a large number of original sample information, but our framework can effectively prevent this. We call the operation of Gaussian blur $Gaussianblur(\cdot)$. Our algorithm is shown in **Algorithm 2**.

4. EXPERIMENT

Dataset: Following previous works (Dong et al., 2018; Lin et al., 2019; Xie et al., 2019), we select the NIPS2017 competition dataset. This dataset extracted 1,000 natural images from the ImageNet dataset and adjusted their size to $299 \times 299 \times 3$.

Network: We selected seven models as our experimental models, including four models under natural training, i.e., Inception-v3 (Inc-v3) (Szegedy et al., 2016), Inception-v4 (Inc-v4) InceptionResnet-v2 (IncRes-v2) (Szegedy et al., 2017), Resnet-v2- 152 (Res-152) (He et al., 2016), and three ensemble adversarial training model (Tramér et al., 2017), i.e., ens3-adv-Inception-v3 (Inc-v3ens3), ens4-adv-Inception-v3 (Inc-v3ens4), and ens-adv-Inception-ResNet-v2 (IncRes-v2ens).

Experimental details: In our experiment, we compare I-FGSM, DI-FGSM, MI-FGSM, SI-FGSM, TI-FGSM, PI-FGSM, and their combined versions, i.e., DI-TI-MI-FGSM, REI-TI-MI-FGSM, and DI-TI-MI-REI-FGSM in the scenario of non-targeted attacks. In our experiment, we set the number of gradient iterations T to 10, the step size α to 1.6, and max perturbation ε to 16. For MI-FGSM, we set the delay factor $\mu = 1.0$; for TI-BIM, we set the kernel size $k = 15$; for DI-FGSM, we set the conversion probability $p = 0.7$; for SI-FGSM, the number of the scale copies m is set to 5; and for PI-FGSM, we set the amplification factor $\beta = 10$.

4.1. The Number and Area of Erasing Matrix

In this section, we discuss the attack performance of the number and area of erasing matrices. Specifically, we choose Inc-v3 as a substitute model to generate adversarial examples and test the results under the other six models with the variable-controlled methods. According to the work by Xie et al. (2021), we set $T = 50$, $a = 1.6$, and $\varepsilon = 16$.

TABLE 3 | The success rate(%) of non-targeted attacks of seven models.

Model	Attacks	Inc-v3	Inc-v4	Res-152	IncRes-v2	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens
Inc-v3	MI-FGSM	**100.0***	55.5	45.3	51.8	22.4	21.0	10.8
	MI-REI-FSGM	99.9	**64.1**	**51.9**	**60.5**	**26.0**	**24.7**	**13.0**
	PI-FGSM	**100.0***	58.6	46.9	50.3	31.4	31.8	20.1
	PI-REI-FGSM	**100.0***	**64.4**	**51.5**	**57.5**	**34.3**	**32.4**	**21.7**
	SI-FGSM	**100.0***	50.5	38.0	44.9	21.6	**21.7**	10.0
	SI-REI-FGSM	99.4*	**78.0**	**65.0**	**74.8**	**44.8**	**45.1**	**26.4**
Inc-v4	MI-FGSM	71.0	**100.0***	51.5	58.4	24.1	23.1	14.0
	MI-REI-FSGM	**78.0**	**100.0***	**57.7**	**65.2**	**28.8**	**27.6**	**16.9**
	PI-FGSM	71.6	**100.0***	50.2	54.4	35.4	35.2	25.0
	PI-REI-FGSM	**76.0**	99.9*	**54.9**	**63.4**	**37.3**	**37.9**	**26.3**
	SI-FGSM	69.9	**100.0***	48.1	55.3	26.9	26.5	14.9
	SI-REI-FGSM	**86.6**	98.9*	**73.2**	**78.5**	**54.0**	**50.5**	**36.1**
Res-152	MI-FGSM	57.5	51.2	**99.2***	47.0	27.1	24.8	15.6
	MI-REI-FSGM	**60.3**	**55.9**	**99.2***	**52.6**	**30.9**	**30.0**	**18.8**
	PI-FGSM	63.6	54.5	**99.7***	50.8	37.5	36.9	26.7
	PI-REI-FGSM	**66.1**	**59.4**	99.3*	**54.8**	**41.0**	**40.4**	**29.4**
	SI-FGSM	43.0	36.3	**99.7***	30.6	20.5	19.2	11.6
	SI-REI-FGSM	**61.8**	**58.1**	97.9*	**54.4**	**40.5**	**38.1**	**27.8**
IncRes-v2	MI-FGSM	77.7	67.0	58	**100.0***	31.6	28.1	20.7
	MI-REI-FSGM	**81.6**	**74.9**	**64.3**	99.7*	**38.4**	**33.9**	**24.3**
	PI-FGSM	76.3	69.4	59.0	**100.0***	40.8	39.1	32.0
	PI-REI-FGSM	**80.6**	**73.9**	**66.1**	99.8*	**45.4**	**43.5**	**36.1**
	SI-FGSM	71.5	58.4	49.8	**100.0***	30.6	28.8	22.5
	SI-REI-FGSM	**84.8**	**80.7**	**76.3**	98.6*	**61.5**	**54.9**	**48.2**

The top row models are substitute models, and we use them to generate adversarial examples by MI-FGSM, PI-FGSM, SI-FGSM, and thier combination with REI-FGSM, ("" indicates the white box attack). The bold value represents the highest success rate for different attack methods under the same experimental conditions.*

4.1.1. Area of Erasing Matrix

Here, we discuss the attack performance under the erasing of a single matrix with different erasing area ratios. As shown in **Figure 4**, with the increase of erasing area, the black-box attack success rate of the three normal models first increases and then remains basically unchanged or slightly decreases, while the attack success rate of the three defense models basically continues to rise. When the erasure area ratio is 0.9, our method can still maintain a high attack success rate, while the attack success rate of the previous framework will decrease very low, indicating the effectiveness of our method. In the normal training model, the attack performance is the best when the erasing area ratio is of 0.5, and in the ensemble adversarial training model, the attack performance is the best when the erasing area ratio is 0.8.

4.1.2. Numer of Erasing Matrix

In this subsection, we discuss the attack performance under different number of erasing matrices with erasing area ratio 0.1. As shown in **Table 1**, with the increase of the number of matrices, the success rate of black-box attack begins to increase. When the number of matrices is 8, the attack on the normal model is the best, and when the number of matrices is 15, the attack on the ensemble adversarial training model is the best. Even if the total erasing area ratio has exceeded 1.0, it can still maintain a high attack success rate, because the initial point of the matrix is randomly selected, and some matrices will overlap so that it does not cover all regions. As shown in **Figure 5**, multiple matrices

TABLE 4 | The success rate(%) of non-targeted attacks of three ensemble adversarial training models.

Model	Attacks	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens
Ensemble	DI-TI-MI-FGSM	94.8	94.5	88.5
	REI-TI-MI-FGSM	94.8	94.5	89.9
	DI-TI-MI-REI-FGSM	**97.6**	**97.3**	**96.2**

The adversarial examples are crafted by DI-TI-MI-FGSM, REI-TI-MI-FGSM, and DI-TI-MI-REI-FGSM on four normal models. The bold value represents the highest success rate for different attack methods under the same experimental conditions.

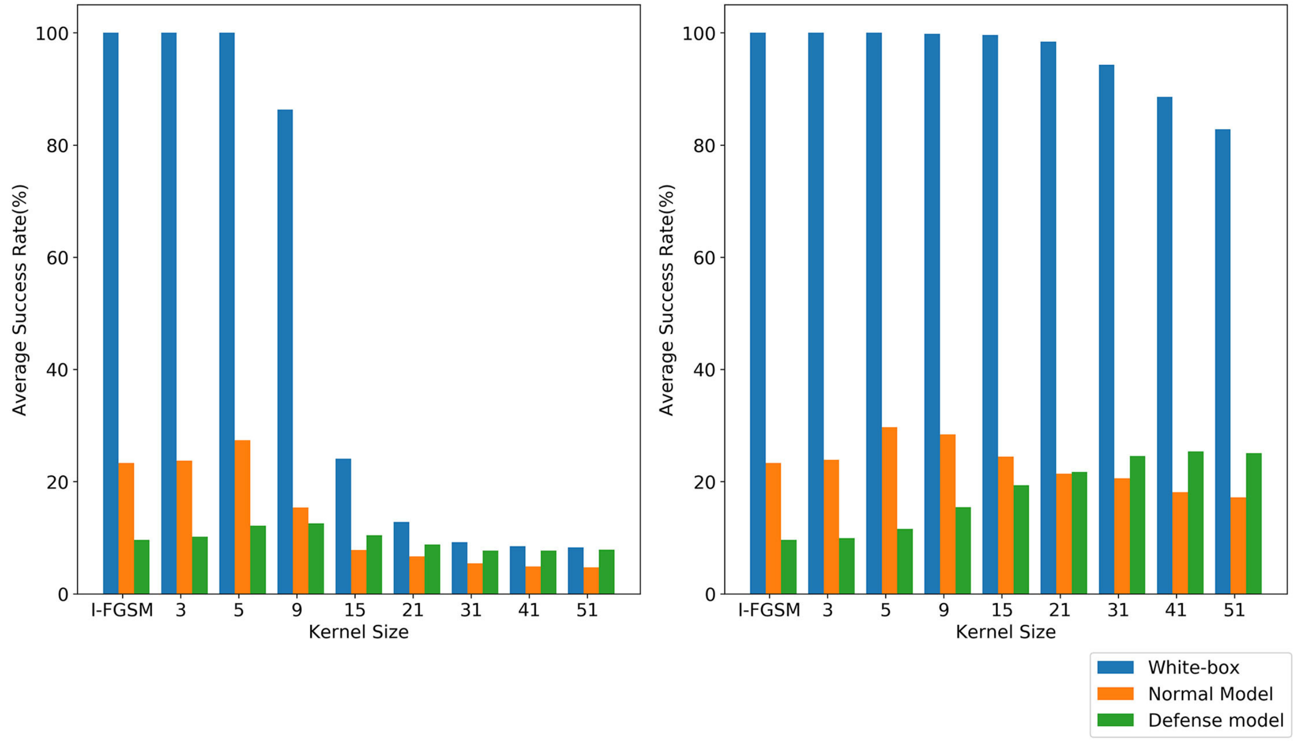


FIGURE 6 | Comparison of GBI-FGSM-F (left) and GBI-FGSM (right).

erasing can transform more shapes than single matrix erasing. We find that when the total area is certain, using more small matrices can achieve better attack results. When the total matrix area is 0.8, the attack success rate of multi-matrix is 2.3% higher than that of a single matrix, and the best attack of multi-matrix is 4.8% higher than that of a single matrix.

4.2. Attack Single Model

In this section, we compare our algorithm with the I-FGSM and data enhancement methods, such as DI-FGSM, SI-FGSM. We also test the experimental results of REI-FGSM combined with MI-FGSM, PI-FGSM and SI-FGSM. The experimental parameters follow the original paper. For REI-FGSM, we set the $\theta_L = \theta_H = 0.1$ and the number of matrices $K = 8$. When combining with PI-FGSM and SI-FGSM, we set $\theta_L = \theta_H = 0.3$ and $K = 3$ for REI-FGSM. When combining with MI-FGSM, we set $\theta_L = \theta_H = 0.1$ and $K = 8$ for REI-FGSM. As shown in **Table 2**, the experimental results show that the attack success rate of our method is 17.3% higher than the I-FGSM on average, 4.2% higher than the DI-FGSM and 2.5% than SI-FGSM. In the defense model, our method is 6.6% higher than DI-FGSM. As shown in **Table 3**, the attack performance of MI-FGSM can be improved by 5.2% on average when combined with REI-FGSM, the attack performance of SI-FGSM can be improved by 22.9% on average when combined with REI-FGSM, and the attack performance of PI-FGSM can be improved by 4.0% on average when combined with REI-FGSM. To sum up, we can find that our method can combine with the above classical methods to achieve greater performance, especially with SI-FGSM, which can increase by an average of 22.9%.

4.3. Attack Ensemble Model

In this section, we use DI-TI-MI-FGSM, REI-TI-MI-FGSM, and DI-TI-MI-REI-FGSM to attack four normal models, and test the success rate of the black-box attack on three ensemble adversarial training models. Following the work (Xie et al., 2021), we set $T = 50$, $a = 3.2$ and $\varepsilon = 16$. For REI-FGSM, we set the $\theta_L = \theta_H = 0.01$ and the number of matrices $K = 30$. As shown in **Table 4**, REI-TI-MI-FGSM achieves an average attack success rate of 93.1% on three defense models, which is 0.5% higher than DI-TI-MI-FGSM. The average attack performance of DI-TI-MI-REI-FGSM can reach 97.0%, which is 4.4% higher than that of DI-TI-MI-FGSM. As far as we know, DI-TI-MI-REI-FGSM achieves the best performance of the current attack method based on gradient iteration.

4.4. Compatibility of the Attack Framework

In order to verify the compatibility of our framework, Gaussian blur (Gedraite and Hadad, 2011) is introduced into our framework. We make use of Gaussian blur attack inc-v3 model in the original framework and our framework, respectively, called GBI-FGSM-F and GBI-FGSM. We take the kernel size as 3,5,9,15,21,31,41, and 51 and compare it with the baseline I-FGSM. As shown in **Figure 6**, with the increase of kernel size, the attack success rate of GBI-FGSM-F decreases significantly, but GBI-FGSM can still maintain a high attack success rate.

Although the attack success rate of GBI-FGSM on the normal model will decrease, the attack success rate on the ensemble adversarial training will increase. We believe that a large degree of disruption for adversarial perturbation during the gradient iteration may result in more robust adversarial examples against defense models. When the kernel size is 51, the attack success rate of GBI-FGSM on the three defense models can reach an average of 25.0%.

5. CONCLUSION

Previous data enhancement frameworks only work on input transformations that satisfy accuracy or loss invariance. However, it does not work for other transformations that do not meet the above conditions, such as the transformation which will lose information. In this paper, we propose a data enhancement framework only for adversarial perturbation, which can effectively solve the above problems. In addition, we introduce random erasing as an input transformation into the generation of adversarial examples for the first time. Compared with the methods based on data enhancement, such as DI-FGSM and SI-FGSM, the attack success rate of REI-FGSM can be improved by 4.2% and 2.5% on average, respectively. DI-TI-MI-REI-FGSM can achieve an average attack success rate of 97.0% on the ensemble adversarial training models, which is better than the current gradient-based iterative method. In addition, we also briefly introduce Gaussian blur to illustrate the compatibility of our framework.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Kaggle, https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack/data.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002502 and National Natural Science Foundation of China (Nos. 61701089, 61601518, and 61372172).

REFERENCES

Behzadan, V., and Munir, A. (2017). "Vulnerability of deep reinforcement learning to policy induction attacks," in *International Conference on Machine Learning and Data Mining in Pattern Recognition* (Cham: Springer), 262–275.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv [Preprint] arXiv:2004.10934.*

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., et al. (2016). End to end learning for self-driving cars. *arXiv [Preprint] arXiv:1604.07316.*

Carlini, N., and Wagner, D. (2017). Towards evaluating the robustness of neural networks. *arXiv [Preprint] arXiv: 1608.04644.* doi: 10.1109/SP.2017.49

Carlini, N., and Wagner, D. (2018). "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)* (San Francisco, CA: IEEE), 1–7.

Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., et al. (2018). "Adversarial attack on graph structured data," in *International Conference on Machine Learning* (Stockholm: PMLR), 1115–1124.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 4690–4699.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., et al. (2018). "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 9185–9193.

Dong, Y., Pang, T., Su, H., and Zhu, J. (2019). "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE), 4312–4321.

Fischer, M., Baader, M., and Vechev, M. (2020). Certified defense to image transformations via randomized smoothing. *arXiv [Preprint] arXiv:2002.12463.*

Gao, L., Cheng, Y., Zhang, Q., Xu, X., and Song, J. (2021). "Feature space targeted attacks by statistic alignment," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization), 671–677.

Gao, L., Zhang, Q., Song, J., Liu, X., and Shen, H. T. (2020). "Patch-wise attack for fooling deep neural network," in *European Conference on Computer Vision* (Cham: Springer), 307–322.

Gedraite, E. S., and Hadad, M. (2011). Investigation on the effect of a gaussian blur in image filtering and segmentation. In Proceedings ELMAR-2011, pages 393-396. IEEE.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). "Convolutional sequence to sequence learning," in *International Conference on Machine Learning* (PMLR), 1243–1252.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv [Preprint] arXiv: 1412.6572.*

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Inkawhich, N., Liang, K. J., Carin, L., and Chen, Y. (2020). Transferable perturbations of deep feature distributions. *arXiv [preprint]* arXiv:2004.12519

Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial examples in the physical world. *arXiv [preprint]* arXiv:1607.02533

Li, Y., Bai, S., Zhou, Y., Xie, C., Zhang, Z., and Yuille, A. (2020). "Learning transferable adversarial examples via ghost networks," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol.* 34, 11458–11465.

Lin, J., Song, C., He, K., Wang, L., and Hopcroft, J. E. (2019). Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv [Preprint] arXiv:1908.06281.*

Liu, W., and Li, Z. (2020). "Enhancing adversarial examples with flip-invariance and brightness-invariance," in *International Conference on Security and Privacy in Digital Economy* (Quzhou: Springer), 469–481.

Liu, Y., Chen, X., Liu, C., and Song, D. (2017). Delving into transferable adversarial examples and black-box attacks. *arXiv[Preprint] arXiv: 1611.02770.*

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks. *arXiv [Preprint] arXiv: 1706.06083.*

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2574–2582.

Papernot, N., McDaniel, P., and Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv [Preprint] arXiv:1605.07277*.

Raghunathan, A., Steinhardt, J., and Liang, P. (2018). Certified defenses against adversarial examples. *arXiv [Preprint] arXiv:1801.09344*.

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv arXiv [Preprint] arXiv:1801.09344*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vo*. 31.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2818–2826.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv [Preprint] arXiv:1312.6199*.

Tramér, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2017). Ensemble adversarial training: attacks and defenses. *arXiv [Preprint] arXiv:1705.07204.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *In Advances in neural information processing systems, pages* 5998–6008.

Wang, X., and He, K. (2021). "Enhancing the transferability of adversarial attacks through variance tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE), 1924–1933.

Wang, X., He, X., Wang, J., and He, K. (2021a). Admix: enhancing the transferability of adversarial attacks. *arXiv [Preprint] arXiv: 2102.00436*. doi: 10.1109/CVPR46437.2021.00196

Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., and Ren, K. (2021b). Feature importance-aware transferable adversarial attacks. *arXiv [Preprint] arXiv: 2107.14185*.

Wu, D., Wang, Y., Xia, S.-T., Bailey, J., and Ma, X. (2020a). Skip connections matter: on the transferability of adversarial examples generated with resnets. *arXiv [Preprint] arXiv: 2002.05990*.

Wu, L., Zhu, Z., Tai, C., and others (2018). Understanding and enhancing the transferability of adversarial examples. *arXiv [Preprint] arXiv:1802.09707*.

Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M. R., et al. (2020b). "Boosting the transferability of adversarial samples via attention," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE), 1158–1167.

Wu, W., Su, Y., Lyu, M. R., and King, I. (2021). "Improving the transferability of adversarial samples with adversarial transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9024–9033.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. (2017). Mitigating adversarial effects through randomization. *arXiv [Preprint] arXiv:1711. 01991*.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., et al. (2019). "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Xie, P., Wang, L., Qin, R., Qiao, K., Shi, S., Hu, G., et al. (2021). Improving the transferability of adversarial examples with new iteration framework and input dropout. *arXiv [Preprint] arXiv:2106.01617*.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020). Random erasing data augmentation. *Proc. AAAI Conf. Artif. Intell*. 34, 13001–13008. doi: 10.1609/aaai.v34i07.7000

Zou, J., Pan, Z., Qiu, J., Liu, X., Rui, T., and Li, W. (2020). "Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting," in *Computer Vision – ECCV 2020, Vol. 12367*, eds A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (Cham: Springer International Publishing), 563–579.

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fnbot.2021.767299

Visual Saliency via Multiscale Analysis in Frequency Domain and Its Applications to Ship Detection in Optical Satellite Images

Ying Yu, Jun Qian and Qinglong Wu*

School of Information Science and Engineering, Yunnan University, Kunming, China

This article proposes a bottom-up visual saliency model that uses the wavelet transform to conduct multiscale analysis and computation in the frequency domain. First, we compute the multiscale magnitude spectra by performing a wavelet transform to decompose the magnitude spectrum of the discrete cosine coefficients of an input image. Next, we obtain multiple saliency maps of different spatial scales through an inverse transformation from the frequency domain to the spatial domain, which utilizes the discrete cosine magnitude spectra after multiscale wavelet decomposition. Then, we employ an evaluation function to automatically select the two best multiscale saliency maps. A final saliency map is generated via an adaptive integration of the two selected multiscale saliency maps. The proposed model is fast, efficient, and can simultaneously detect salient regions or objects of different sizes. It outperforms state-of-the-art bottom-up saliency approaches in the experiments of psychophysical consistency, eye fixation prediction, and saliency detection for natural images. In addition, the proposed model is applied to automatic ship detection in optical satellite images. Ship detection tests on satellite data of visual optical spectrum not only demonstrate our saliency model's effectiveness in detecting small and large salient targets but also verify its robustness against various sea background disturbances.

Keywords: visual saliency, selective visual attention, wavelet transform, multiscale saliency map, ship detection

OPEN ACCESS

Edited by:
Jingyu Hou,
Deakin University, Australia

Reviewed by:
Nemir Al-Azzawi,
University of Baghdad, Iraq
Deepika Koundal,
University of Petroleum and Energy Studies, India

****Correspondence:***
Ying Yu
yuying.mail@163.com

Received: *30 August 2021*
Accepted: *01 December 2021*
Published: *13 January 2022*

Citation:
Yu Y, Qian J and Wu Q (2022) Visual Saliency via Multiscale Analysis in Frequency Domain and Its Applications to Ship Detection in Optical Satellite Images.
Front. Neurorobot. 15:767299.
doi: 10.3389/fnbot.2021.767299

INTRODUCTION

In the human neural system, a mechanism called selective visual attention has been evolved to facilitate our visual perception to rapidly locate the most important regions in a cluttered scene. Such important regions are said to be perceptually salient because they attract great visual attention. Typically, visual attention is either driven by fast, pre-attentive, bottom-up visual saliency or controlled by slow, task-dependent, top-down cues (Itti et al., 1998; Itti and Koch, 2001; Wolfe and Horowitz, 2004, 2017).

This article is primarily concerned with the automatic detection of bottom-up visual saliency, which has attracted extensive studies by both psychologist and computer vision researchers in the area of robotics, cognitive science, and neuroscience (Borji and Itti, 2013). Just like a bottom-up visual attention mechanism that can rapidly locate salient objects in the human visual pathway, a computational saliency model has the ability to detect the perceptually salient regions in cluttered scenes, which is very useful for object detection, image segmentation, intelligent compression, human fixation prediction, and many more.

One pioneer work concerning the computational modeling of bottom-up visual attention was introduced by Itti et al. (1998) and Itti and Koch (2001). It stimulates the neural mechanism of the human early vision system and has explicit biological rationality. Itti's model (denoted IT) generates a saliency map of the scene under view by modeling the center-surround contrast of intensity, color, and orientation, which is expected to indicate salient regions and predict human fixations. However, since the model is designed conforming to the neuronal architecture of a vision system, it is computationally complex and suffers from over-parameterization. Recently, a kind of algorithm has been designed for salient foreground segmentation. Achanta and Suesstrunk (2010) compute saliency maps by use of the Euclidean distance in the Commission International Eclairage (CIE) LAB space between a given position's value and the maximum symmetric surround mean value of its neighboring area (denoted MSS). Cheng et al. (2015) used a histogram-based contrast (HC) to measure the saliency values of input images. Liu and Yang (2019) exploited color volume and color difference for salient region detection. These algorithms can output fine-resolution saliency maps that highlight large-scale foreground regions. However, since these kinds of algorithms are not biologically motivated, they cannot be used as a model of bottom-up visual attention with psychophysical consistency and often fail to detect salient objects in cluttered scenes.

Another kind of bottom-up saliency model is computed in the frequency domain. These frequency-domain models are not explicitly motivated by a biological mechanism, but they are computationally simple and have good consistency with psychophysics. As a pioneer saliency work of frequency domain, Hou and Zhang (2007) designed a saliency model by use of the Fourier spectral residual (SR) computation. Yu et al. (2009) proposed the pulsed functions of the discrete cosine transform (PCT) to compute visual saliency. Guo and Zhang (2010) introduced a spatiotemporal saliency approach by using a so-called phase spectrum of quaternion Fourier transform (PQFT). Li et al. (2013) compute visual saliency via a scale-space analysis in the hypercomplex Fourier transform (HFT) domain. After that, Yu and Yang (2017) proposed a visual saliency model by using the binary spectrum of Walsh–Hadamard transform (BSWHT).

As for why the frequency domain models can calculate visual saliency, our previous works (Yu et al., 2011a,b) have demonstrated the biological rationality of frequency-domain approaches. These works have verified that whitening or flattening the principal components or the cosine transform coefficients simulates the suppression of the same visual features (iso-feature suppression) in the spatial domain. The iso-feature suppression is just the biological mechanism of bottom-up visual saliency generated in the primary visual cortex (V1) (Zhaoping, 2002; Zhaoping and Peter, 2006). However, due to the excessive suppression of low-frequency components in the image by whitening the principal components, existing frequency-domain models are easy to detect small salient targets, but they have poor ability to highlight large-scale salient regions.

To make the frequency domain model have better detection ability for both large and small salient targets, in this article, we propose a bottom-up visual saliency model based on multiscale analysis and computation in the frequency domain. The proposed model performs multiscale wavelet analysis and computation in the cosine transform domain. It can generate multiscale saliency maps of the scene under view. Unlike the spatial domain approaches, our model computes in the frequency domain, which significantly reduces computational cost for a saliency algorithm. Moreover, the multiscale computation of visual saliency also has biological plausibility because the receptive fields of visual neurons in the primary visual cortex (V1) have various ranges of center-surround mechanism (Itti and Koch, 2001; Zhaoping, 2002; Zhaoping and Peter, 2006). As compared with the existing frequency domain approaches, our model has a better ability to detect small salient objects and meanwhile highlight large-scale salient regions.

Ship target detection in optical satellite images is important in monitoring commercial fishery, oil pollution, vessels traffic, and other marine activities. However, there remain challenges with the ship detection algorithm for its application in a marine surveillance system. One challenge is the existence of sea clutters and heterogeneous regions, which poses difficulties for discriminating ship targets from various background disturbances. Another challenge is that a marine surveillance system needs fast algorithms because it needs to analyze and process large amounts of data in real-time. In this work, we apply our multiscale saliency model to detect the ship signatures in the optical satellite images. It may meet the demands of a marine surveillance system and can detect ships of different sizes accurately. Tests over the Maritime SATellite Imagery (MASATI) dataset prove the robustness and effectiveness of our model when it is applied to ship detection in optical satellite images.

The rest of this article is organized as follows. Section Proposed Model describes the proposed bottom-up visual saliency model based on multiscale analysis and computation in the frequency domain and explains its biological plausibility. Section Experimental Validation presents our model's experiments on psychophysical patterns, eye fixation prediction, and saliency detection for natural images. In section Applications to Ship Detection in Optical Satellite Images, we apply the proposed saliency model to automatic ship detection in optical satellite images. Finally, this article is concluded in section Conclusion and Discussion.

PROPOSED MODEL

This section begins by introducing the proposed model of bottom-up visual saliency step by step, and then gives a complete flow of the model from the input image to a final saliency map.

Visual Feature Channels

Several works (Treisman and Gelade, 1980; Zhaoping, 2002; Zhaoping and Peter, 2006) have verified that the interaction and integration of the low-level visual features can produce a bottom-up saliency map in the primary visual cortex (V1). To begin with, we will compute these low-level visual feature maps before integrating them as a whole. For a given image $\boldsymbol{M}$ (e.g., resized to 128 × 128 px), we use $\boldsymbol{r}$, $\boldsymbol{g}$, and $\boldsymbol{b}$ to denote the red, green, and blue color channels of the image, respectively. According to Itti et al.'s (1998) general-tuned color model, one intensity and three general-tuned color feature channels $\boldsymbol{I}$, $\boldsymbol{R}$, $\boldsymbol{G}$, and $\boldsymbol{B}$ are calculated as

$$\boldsymbol{I} = \frac{\boldsymbol{r}+\boldsymbol{g}+\boldsymbol{b}}{3} \tag{1a}$$

$$\boldsymbol{R} = \boldsymbol{r} - \frac{\boldsymbol{g}+\boldsymbol{b}}{2} \tag{1b}$$

$$\boldsymbol{G} = \boldsymbol{g} - \frac{\boldsymbol{r}+\boldsymbol{b}}{2} \tag{1c}$$

$$\boldsymbol{B} = \boldsymbol{b} - \frac{\boldsymbol{r}+\boldsymbol{g}}{2} \tag{1d}$$

Note that the general-tuned red, green, and blue channels $\boldsymbol{R}$, $\boldsymbol{G}$, and $\boldsymbol{B}$ are set to zero at locations with a negative value.

In the primary visual cortex (V1) of the human brain, similar neurons have lateral inhibition, that is, excited neurons will inhibit the surrounding similar neurons so that the unique targets in the visual scene are highlighted and become the salient targets obtained by the visual attention mechanism (Zhaoping and Peter, 2006). Referring to the lateral inhibition process, we can consider that a red flower in the green grass is salient. If the color feature energy is considered as the sum of the pixels of the color feature channel, then the green feature channel has the largest energy in the scene. Conforming to the characteristics of selective visual attention, our model adjusts the weight of each color feature channel to reduce the weight factor of the feature channel with large energy. In this article, the weight factors of each feature channel in the visual saliency map are defined as

$$\begin{cases} \omega_{\mathcal{M}} = \frac{\max(\mathcal{M})}{\sqrt{\sum_{i=1}^{128}\sum_{j=1}^{128}\mathcal{M}}}, & if \quad \sum_{i=1}^{128}\sum_{j=1}^{128}\mathcal{M} \neq 0 \\ \omega_{\mathcal{M}} = \max(\mathcal{M}), & if \quad \sum_{i=1}^{128}\sum_{j=1}^{128}\mathcal{M} = 0 \end{cases} \tag{2}$$

where $\mathcal{M}$ denotes any one of the general-tuned feature channels $\boldsymbol{I}$, $\boldsymbol{R}$, $\boldsymbol{G}$, and $\boldsymbol{B}$, whereas i and j are the horizontal and vertical coordinates of the corresponding channel.

Multiscale Saliency Computation in the Frequency Domain

After calculating the visual feature channels of the input image, we perform multiscale saliency computation and analysis in the frequency domain. Given a visual feature channel $\mathcal{M}$, we use the discrete cosine transform (DCT) to transform each visual feature channel of the image into a frequency domain:

$$\boldsymbol{\mathcal{F}} = \mathrm{DCT}(\mathcal{M}) \tag{3}$$

where "DCT(·)" denotes a 2-dimensional discrete cosine transform, and $\boldsymbol{\mathcal{F}}$ is the DCT coefficients matrix of the input visual feature channel. Next, the magnitude matrix $\boldsymbol{A}_{\mathcal{M}}$ and the sign matrix $\boldsymbol{S}_{\mathcal{M}}$ of the DCT coefficients matrix $\boldsymbol{\mathcal{F}}$ are computed as

$$\begin{cases} \boldsymbol{A}_{\mathcal{M}} = \mathrm{abs}(\boldsymbol{\mathcal{F}}) \\ \boldsymbol{S}_{\mathcal{M}} = \mathrm{sign}(\boldsymbol{\mathcal{F}}) \end{cases} \tag{4}$$

where the notation "abs(·)" is an absolute value function, and the notation "sign(·)" denotes a signum function. For most input images, the magnitude values of low-frequency coefficients are much greater than those of high-frequency coefficients since the natural images have a strong statistical correlation in the visual space. Our previous works (Yu et al., 2011a,b) have verified that whitening or flattening the principal components or the cosine transform coefficients simulates the suppression of the same visual features (iso-feature suppression) in the spatial domain. The iso-feature suppression is just the biological mechanism of bottom-up visual saliency generated in the primary visual cortex (V1) (Zhaoping, 2002; Zhaoping and Peter, 2006). Most frequency domain-based models (e.g., Yu et al., 2009, 2011a,b; Guo and Zhang, 2010; Yu and Yang, 2017) can detect relatively small salient objects by setting the values of the magnitude matrix to one. For salient objects with very large sizes, they often highlight the contour of a large object because whitening (flattening) the magnitude matrix will lose some important low-frequency information.

To make the frequency domain model have better detection ability for both large and small salient targets, in this work, we propose a bottom-up visual saliency model based on multiscale analysis and computation in the frequency domain. The proposed model not only detect small salient objects but also highlight the whole body of those salient objects with very large size. We consider utilizing the wavelet transform to perform multiscale modulation on the magnitude matrix of the DCT coefficients.

Wavelet transform is widely used in image decomposition and reconstruction, which can decompose an image into multiscale components. In this article, we employ wavelet transform to decompose the magnitude matrix of each visual feature channel and suppress the low-frequency components of the magnitude matrix to a certain extent. Since the salient targets have different sizes in the image, the retention degree of the values in the required magnitude matrix is different. Therefore, we perform multiscale decomposition and reconstruction of the magnitude matrix of each feature channel, and construct a multiscale reconstruction magnitude matrix set $\left\{\boldsymbol{A}'_{\mathcal{M},\mathcal{N}}\right\}$, where $\mathcal{M}$ is the feature channel set, and $\mathcal{N}$ denotes the decomposition scale. This process ensures that the optimal reconstructed magnitude matrix of the input image can be retained. In the j-scale space, the Mallat decomposition formula of the low-frequency subband is

$$\begin{cases} \boldsymbol{B}_{\mathrm{ss}}{}_{i,l}^{j-1} = \sum\limits_{k,m} \boldsymbol{h}(k-2i)\boldsymbol{h}(m-2l)\boldsymbol{B}_{\mathrm{ss}}{}_{k,m}^{j} \\ \boldsymbol{B}_{\mathrm{ds}}{}_{i,l}^{j-1} = \sum\limits_{k,m} \boldsymbol{g}(k-2i)\boldsymbol{h}(m-2l)\boldsymbol{B}_{\mathrm{ss}}{}_{k,m}^{j} \\ \boldsymbol{B}_{\mathrm{sd}}{}_{i,l}^{j-1} = \sum\limits_{k,m} \boldsymbol{h}(k-2i)\boldsymbol{g}(m-2l)\boldsymbol{B}_{\mathrm{ss}}{}_{k,m}^{j} \\ \boldsymbol{B}_{\mathrm{dd}}{}_{i,l}^{j-1} = \sum\limits_{k,m} \boldsymbol{g}(k-2i)\boldsymbol{g}(m-2l)\boldsymbol{B}_{\mathrm{ss}}{}_{k,m}^{j} \end{cases} \tag{5}$$

and the corresponding reconstruction formula is

$$\begin{aligned} \boldsymbol{B}_{\mathrm{ss}}{}_{k,m}^{j} = & \sum_{i,l} [\boldsymbol{B}_{\mathrm{ss}}{}_{i,l}^{j-1}\boldsymbol{h}(k-2i)\boldsymbol{h}(m-2l) + \boldsymbol{B}_{\mathrm{ds}}{}_{i,l}^{j-1}\boldsymbol{g}(k-2i)\boldsymbol{h}(m-2l) \\ & + \boldsymbol{B}_{\mathrm{sd}}{}_{i,l}^{j-1}\boldsymbol{h}(k-2i)\boldsymbol{g}(m-2l) + \boldsymbol{B}_{\mathrm{dd}}{}_{i,l}^{j-1}\boldsymbol{g}(k-2i)\boldsymbol{g}(m-2l)] \end{aligned} \tag{6}$$

where $\boldsymbol{h}$ and $\boldsymbol{g}$ denote low-pass and high-pass filtering, respectively. As has been noted before, the suppression of cosine transform coefficients of an image is equivalent to the suppression of the same visual features (iso-feature suppression) in the spatial domain. Therefore, through such a multiscale decomposition and reconstruction operation upon the magnitude coefficients in the DCT domain, our model simulates the cortical center-surround or iso-feature suppression of various scales in the spatial domain. For this reason, our model can compute the multiscale saliency information simultaneously, which is very helpful to detect salient objects of different sizes.

To recover the multiscale channel conspicuity maps in the visual space, we perform an inverse DCT on the reconstructed magnitude matrix and the corresponding sign matrix $\boldsymbol{S}_{\mathcal{M}}$ as

$$\boldsymbol{F}_{\mathcal{M},\mathcal{N}} = \text{abs}(\text{IDCT}(\boldsymbol{S}_{\mathcal{M}} \cdot \boldsymbol{A}'_{\mathcal{M},\mathcal{N}})) \tag{7}$$

where $\boldsymbol{F}_{\mathcal{M},\mathcal{N}}$ denotes the $\mathcal{N}$-scale conspicuity map for a given channel $\mathcal{M}$, and "IDCT$(\cdot)$" is the inverse discrete cosine transform. Afterward, we utilize the obtained one intensity and three-color conspicuity maps at the $\mathcal{N}$-scale to compute the saliency map:

$$\boldsymbol{S}_{\mathcal{N}} = \Phi * (\omega_I \cdot \boldsymbol{F}_{I,\mathcal{N}} + \omega_R \cdot \boldsymbol{F}_{R,\mathcal{N}} + \omega_G \cdot \boldsymbol{F}_{G,\mathcal{N}} + \omega_B \cdot \boldsymbol{F}_{B,\mathcal{N}}) \tag{8}$$

where ω_I, ω_R, ω_G, and ω_B denote the weight factors of corresponding feature channels $\boldsymbol{I}$, $\boldsymbol{R}$, $\boldsymbol{G}$, and $\boldsymbol{B}$, which are calculated by using Equation (2). The notation Φ denotes a 2-dimensional Gaussian low-pass filter. The notation $\boldsymbol{S}_{\mathcal{N}}$ is the $\mathcal{N}$-scale saliency map of the input image.

Final Saliency Map

To generate the optimal visual saliency map from the multiscale saliency maps $\{\boldsymbol{S}_{\mathcal{N}}\}$, we introduce an evaluation function to evaluate the multiscale saliency maps. More often than not, the more complete the salient region in multiple saliency maps of the same scene, the better the saliency map with less background interference. The evaluation function is defined as the noise coefficient of the saliency map multiplied by the information entropy, where the noise coefficient is the sum of the product of the pixels corresponding to the background interference matrix and the saliency map matrix. According to the visual attention characteristics that the central area of the image is more likely to become the salient region, the background interference matrix is constructed as a gradient matrix with the same resolution as the saliency map, with a maximum value of 1 and a minimum value of 0. Information entropy is often used as the quantitative standard for evaluating images. In this work, we use information entropy to characterize the degree of confusion of a saliency map. The greater the entropy, the more chaotic the saliency map, that is, the more background interference. Therefore, for a given saliency map S, the corresponding evaluation function H can be defined as

$$H = E \sum_x \sum_y \boldsymbol{S}(x, y)\boldsymbol{K}(x, y) \tag{9}$$

where E is the information entropy of the saliency map $\boldsymbol{S}$, and $\boldsymbol{K}$ denotes the background interference matrix. x, y are the horizontal and vertical coordinates of a matrix. According to the definition of the evaluation function, the smaller the function value, the better the saliency map.

To improve the adaptability of the model in this article, two saliency maps with the lowest value of the evaluation function are selected. Next, the evaluation function values of the corresponding saliency map are exchanged as coefficients to construct the fusion map, and the fusion map is used as the final saliency map $\boldsymbol{\S}$ after central bias optimization. This calculation process is formulated as

$$\boldsymbol{\S} = \boldsymbol{\psi} \cdot (H_2\boldsymbol{S}_1 + H_1\boldsymbol{S}_2) \tag{10}$$

where $\boldsymbol{\psi}$ is the central bias matrix. $\boldsymbol{S}_1$ denotes the saliency map with the smallest evaluation function value, whereas $\boldsymbol{H}_1$ is the corresponding evaluation function value of $\boldsymbol{S}_1$. When there is little difference in the values of the evaluation function, the two saliency maps generate the final saliency map close to their mean value. When the difference of H_1 and H_2 is large, $\boldsymbol{S}_2$ has a weak effect on the generation of the final saliency map.

To sum up, the proposed computational model from input image $\boldsymbol{M}$ to final saliency map $\boldsymbol{\S}$ is as follows:

Step 1. Compute one intensity and three general-tuned color feature channels $\boldsymbol{I}$, $\boldsymbol{R}$, $\boldsymbol{G}$, and $\boldsymbol{B}$ by using Equation (1), and calculate the weight factor $\omega_{\mathcal{M}}$ for each feature channel by using Equation (2).

Step 2. Perform a DCT transformation on each feature channel, and calculate the magnitude matrix $\boldsymbol{A}_{\mathcal{M}}$ and the sign matrix $\boldsymbol{S}_{\mathcal{M}}$ of the DCT coefficients by using Equations (3) and (4).

Step 3. Perform multiscale wavelet transform on all feature channels to obtain the multiscale reconstruction magnitude matrix set $\left\{\boldsymbol{A}'_{\mathcal{M},\mathcal{N}}\right\}$ by using Equations (5) and (6).

Step 4. Performing an inverse DCT transformation on the magnitude matrix $\boldsymbol{A}'_{\mathcal{M},\mathcal{N}}$ and the corresponding sign matrix $\boldsymbol{S}_{\mathcal{M}}$ to compute the $\mathcal{N}$-scale conspicuity map $\boldsymbol{F}_{\mathcal{M},\mathcal{N}}$ by using Equation (7).

Step 5. Performing a weighted summation of the conspicuity map of all four feature channels to compute the $\mathcal{N}$-scale saliency map $S_{\mathcal{N}}$ by using Equation (8).

Step 6. Compute the evaluation function value of the $\mathcal{N}$-scale saliency map $\boldsymbol{S}_{\mathcal{N}}$ by using Equation (9).

Step 7. The two saliency maps with the smallest value of the evaluation function are selected to generate a final saliency map $\boldsymbol{\S}$ by using Equation (10).

The complete flow of the proposed model is illustrated in **Figure 1**. We initially resize the input image to a suitable scale and decompose it into the general-tuned intensity, red, green, and blue feature channels. Each of the four general-tuned feature channels is subjected to a DCT. Next, we use a multiscale wavelet transform to decompose the DCT magnitude spectrum of each channel and then obtain the decomposed multiscale magnitude spectra for every single channel. Afterward, the decomposed magnitude coefficients are subjected to an inverse DCT so that the six multiscale conspicuity maps of each feature channel can be generated. Then, for each scale, we integrate the four conspicuity maps to form a saliency map. Finally, a final saliency

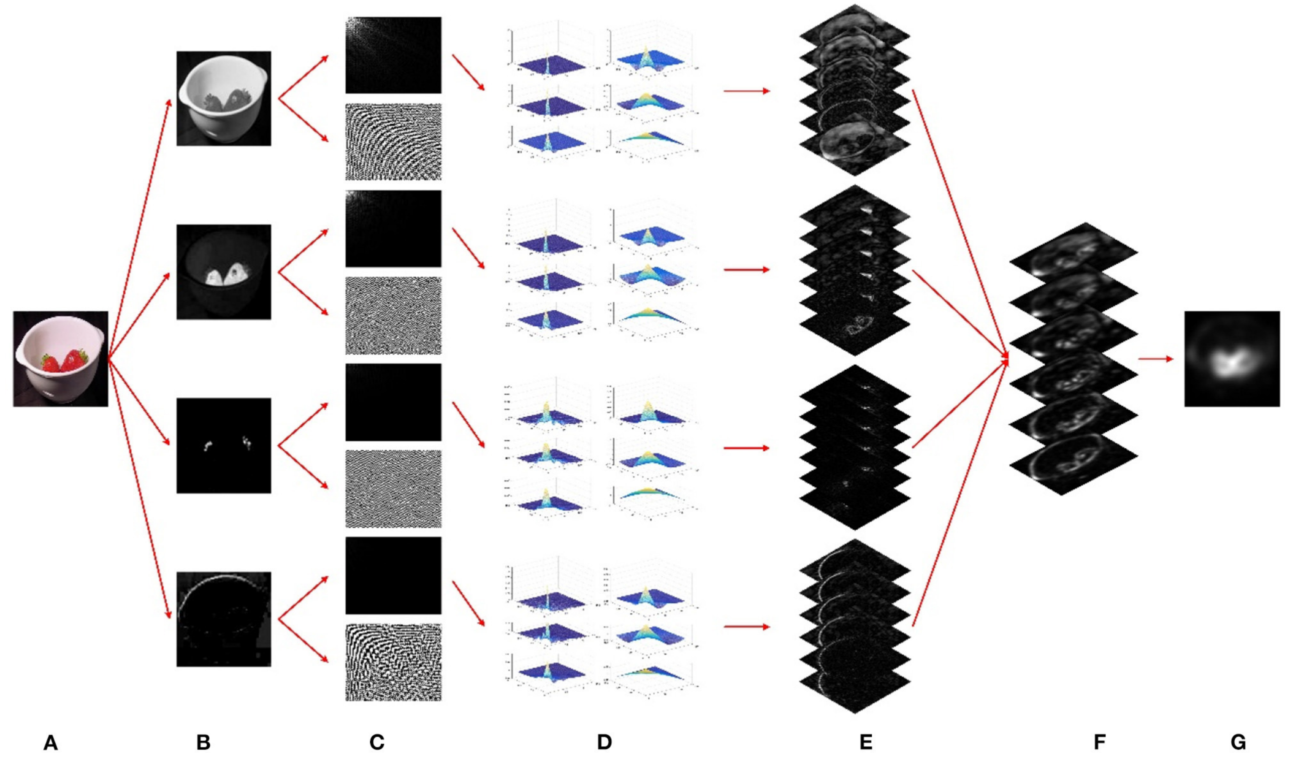


FIGURE 1 | An illustration of our model for saliency detection. **(A)** Input image. **(B)** Four general-tuned visual channels. **(C)** Discrete cosine transform (DCT) magnitude spectra and corresponding signum spectra. **(D)** Wavelet decomposed multiscale magnitude spectra for every single channel. **(E)** Six multiscale conspicuity maps for every single channel. **(F)** Six multiscale saliency maps after channel-wise integration. **(G)** A final saliency map after scale-wise combination.

map is obtained by combining the two multiscale saliency maps with the smallest H-value. Note that the saliency map is a topographically arranged map that represents the visual saliency of a corresponding visual scene. It can be seen from **Figure 1** that the salient objects are the strawberries, which pop out from the background in the final saliency map.

It is worth noting again that the flattening modulation of image frequency domain coefficients approximately simulates the suppression of the same visual features (iso-feature suppression) in the spatial domain. Such a mechanism of iso-feature suppression generates bottom-up visual saliency in the primary visual cortex (V1). In this work, we employ multiscale frequency domain modulation by using a multiscale wavelet transform on the magnitude coefficients in the DCT domain. This calculation process is equivalent to flattening the frequency domain coefficients in different degrees (see **Figure 1D**), rather than in a single way, to calculate the multiscale visual saliency (see **Figure 1F**) in the spatial domain.

EXPERIMENTAL VALIDATION

In this section, we compare our model with eight bottom-up saliency approaches: Itti's model (IT) (Itti et al., 1998), maximum symmetric surround mean value (MSS) (Achanta and Suesstrunk, 2010), histogram-based contrast (HC) (Cheng et al., 2015), spectral residual (SR) (Hou and Zhang, 2007), pulsed cosine transform (PCT) (Yu et al., 2009), PQFT (Guo and Zhang, 2010), hypercomplex Fourier transform (HFT) (Li et al., 2013), and binary spectrum of Walsh-Hadamard transform (BSWHT) (Yu and Yang, 2017). All saliency approaches are conducted on psychophysical pattern tests, human eye fixation prediction, and saliency detection for natural images. The experiments provide an objective evaluation as well as a visual comparison of all saliency maps. Moreover, we give a comparison of the computational time cost of all saliency approaches.

In the experiments of human eye fixation prediction and natural image saliency detection. We will employ three popular objective evaluation metrics: the precision-recall (P-R) curve (Davis and Goadrich, 2006), the receiver operating characteristic (ROC) curve (Tatler et al., 2005), and the area under the curve (AUC). For each saliency map, several binary maps are generated by segmenting the saliency map with a threshold τ varying from 0 to 255. We can obtain the true positive (TP), the false positive (FP), the false negative (FN), and the true negative (TN) by comparing a binary map with the ground truth (GT) map. Then, the *Recall* and the *Precision* metrics for a binary map can be calculated as

$$\begin{cases} Recall = \frac{TP}{TP+FN} \\ Precision = \frac{TP}{TP+FP} \end{cases} \quad (11)$$

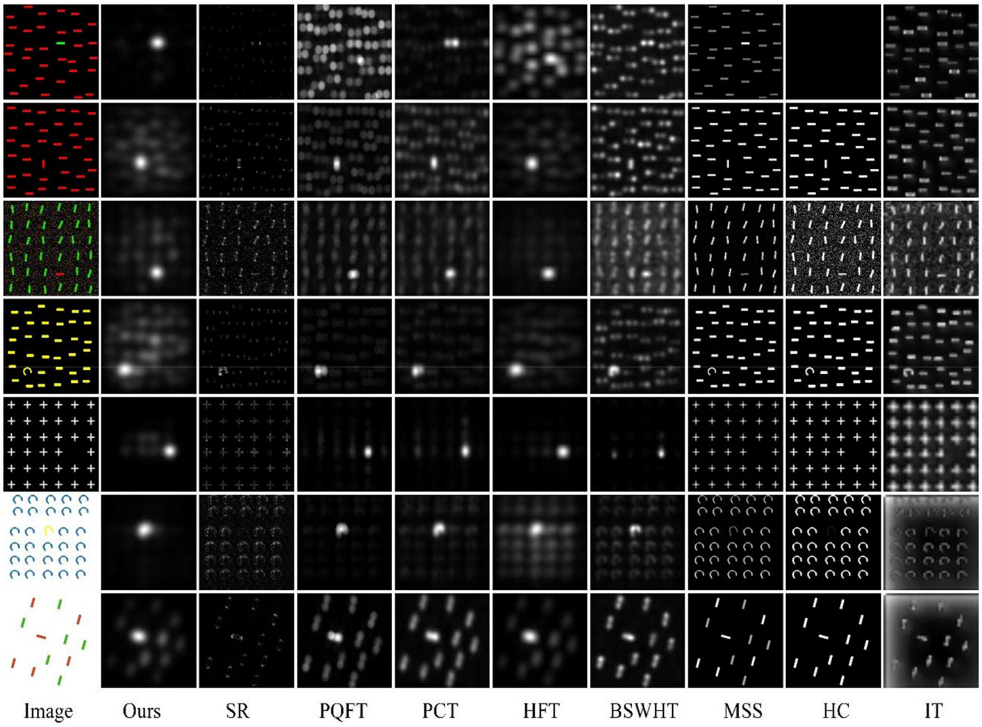


FIGURE 2 | Test on psychophysical patterns for all 9 saliency approaches.

TABLE 1 | The receiver operating characteristic (ROC)- area under the curve (AUC) scores of all nine saliency methods.

Method	Ours	SR	PQFT	PCT	HFT	BSWHT	MSS	HC	IT
All fixations	0.7889	0.6228	0.7570	0.7605	0.7653	0.7761	0.6558	0.5766	0.5365
First fixations	0.8252	0.6274	0.7696	0.7723	0.7902	0.7913	0.6698	0.5850	0.5444

The P-R curve can be plotted with the averaged *Precision* vs. *Recall* values overall saliency maps generated from a saliency approach. Moreover, we compute the true positive rate (*TPR*) and the false positive rate (*FPR*) according to the following formulas:

$$\begin{cases} TPR = \frac{TP}{TP+FN} \\ FPR = \frac{FP}{FP+TN} \end{cases} \tag{12}$$

The ROC curve can be plotted with the averaged *TPR* vs. *FPR* values overall saliency maps generated from a saliency approach. Then we compute the area under the ROC curve that is denoted as a ROC-AUC score. Note that most published articles use these three metrics to evaluate a saliency map's ability to predict eye fixations or detect salient regions.

Psychophysical Consistency

Psychophysical patterns have been widely used in attention selection tests not only to explore the mechanism of bottom-up attention but also to evaluate the saliency models (e.g., Itti et al., 1998; Hou and Zhang, 2007; Yu et al., 2009, 2011a,b; Guo and Zhang, 2010; Li et al., 2013). **Figure 2** shows the saliency maps of all saliency approaches on seven psychophysical patterns (including salient targets of unique color, orientation, shape, missing feature, or conjunction feature). It can be seen that IT, MSS, and HC fail to detect (highlight) the salient targets with distinctive orientation or shape. SR cannot detect color saliency since it only computes in an intensity channel. As frequency-domain approaches, PQFT, PCT, HFT, BSWHT, and our method (denoted as "Ours") can successfully detect salient objects with distinctive orientation or missing features (the 5th pattern). It can be noticed that PCT and our method can find all salient objects with distinctive colors; whereas PQFT and HFT cannot highlight the color pop-out in the 1st pattern. In this test, PCT and our method are the best performers, which are highly consistent with human perception in these psychophysical patterns.

It is worth stating that this article proposes a visual saliency method based on frequency domain calculation. At present, all frequency-domain visual saliency methods do not calculate pixel by pixel, and the output saliency map does not have a clear and accurate object contour. Before outputting the final saliency map, these frequency-domain methods need to do low-pass filtering to obtain an applicable and smooth visual saliency map.

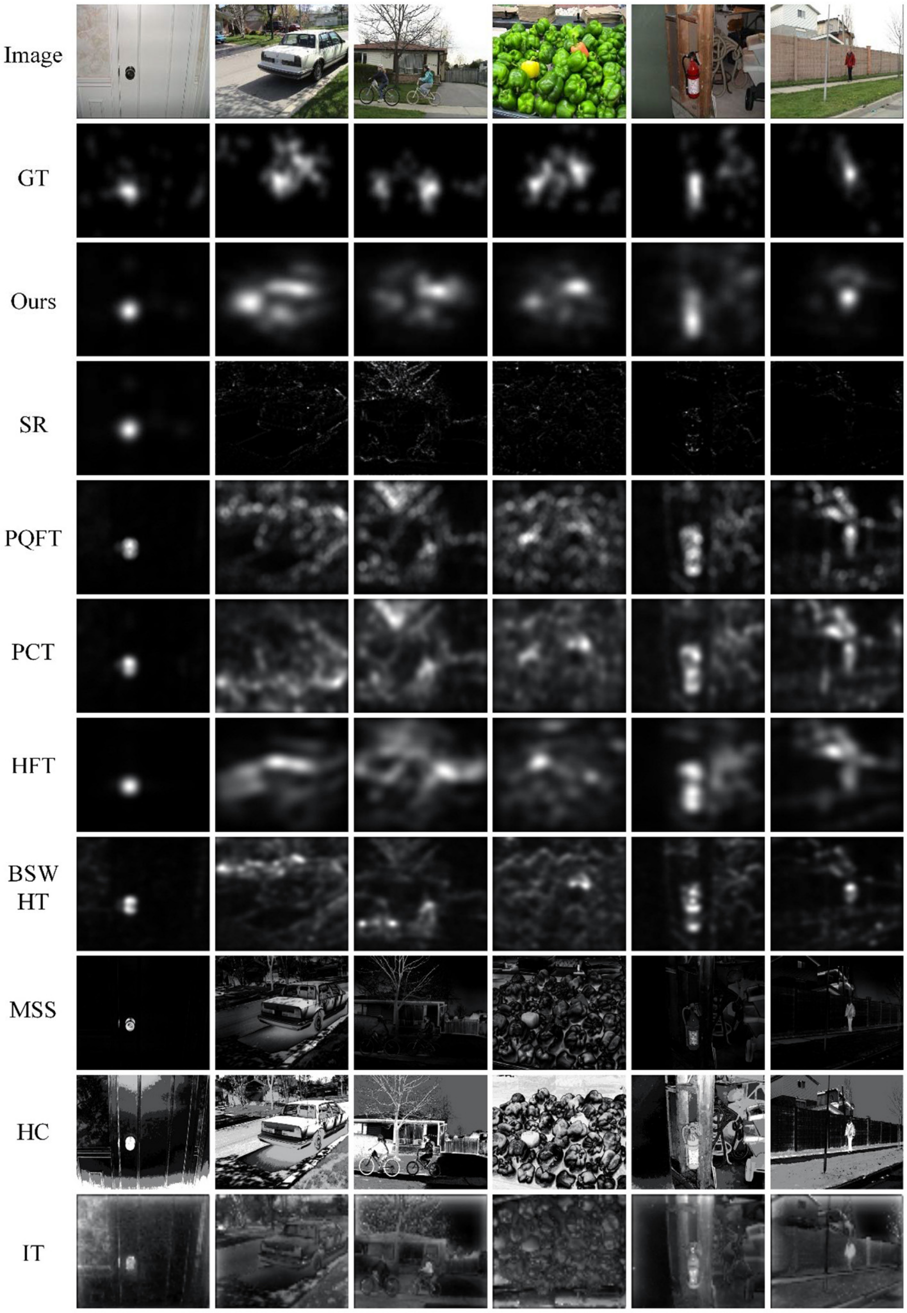


FIGURE 3 | Qualitative analysis of the saliency maps for eye fixation prediction.

Nevertheless, the advantages of frequency-domain methods are also obvious. They can indicate the salient positions and regions in the visual scene, and can better predict the gaze or fixations driven by a human's bottom-up attention mechanism.

Eye Fixation Prediction

In this subsection, we validate the proposed saliency maps by use of the dataset of 120 color images from an urban environment and corresponding human eye fixation data from 20 subjects provided by Bruce and Tsotsos (2009). These images consist of indoor and outdoor scenes, of which some have very salient items, and others have no particular regions of salience.

To quantify the consistency of a particular saliency map with a set of fixations of the image, wey employ the ROC-AUC score as an objective evaluation metric. It is worth noting that the ROC-AUC score is sensitive to the number of fixations that are used in the calculation. Former fixations are more likely to be driven by the bottom-up manner, whereas later fixations are more likely to be influenced by top-down cues. In this test, we calculate the ROC-AUC scores for each image by using all fixations and

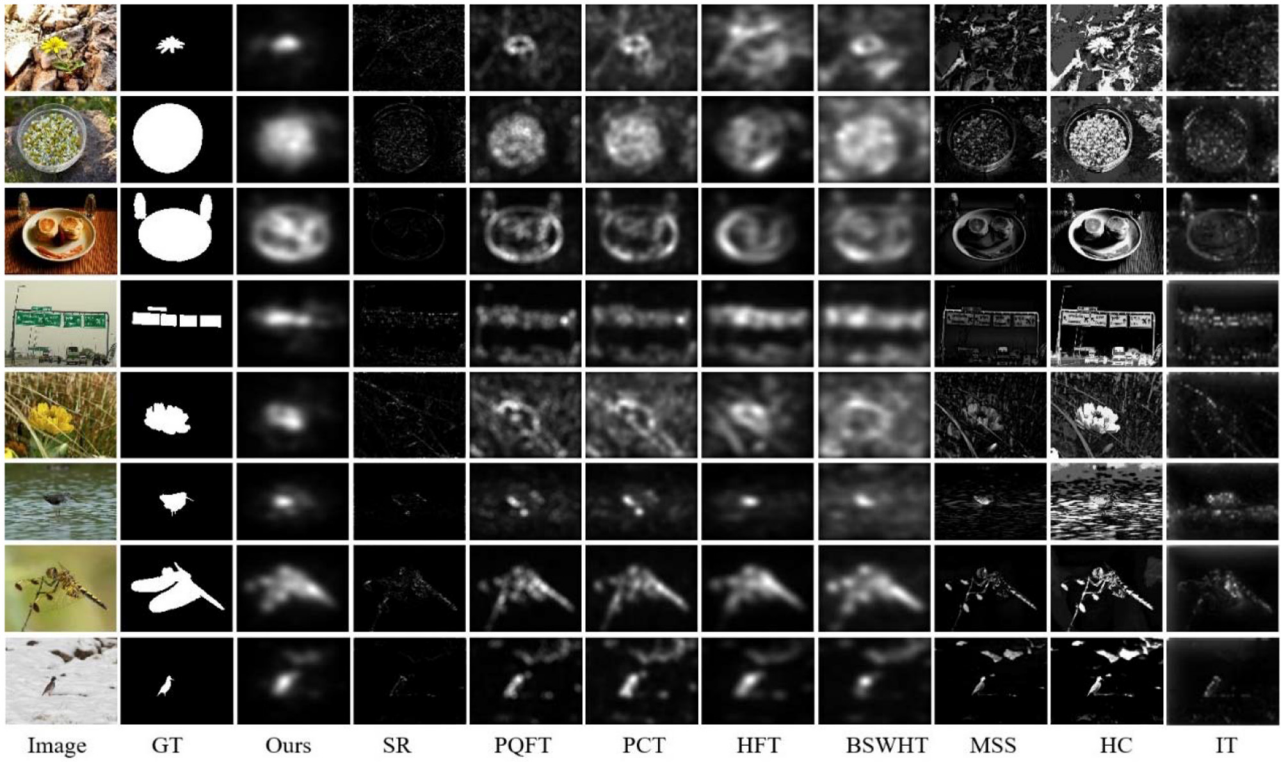


FIGURE 4 | Visual comparison of all saliency approaches on ECSSD dataset.

repeating the process but using only the first two fixation points. **Table 1** lists the ROC-AUC score averaged over all 120 images for each saliency approach. As can be seen, our method obtains the highest ROC-AUC scores in both tests and therefore has the best capability for predicting eye fixations.

Figure 3 gives the saliency maps for six representative images from the data set, which provides a qualitative comparison of all saliency methods. We generate corresponding ground truth images by using a Gaussian filter to perform convolution on the fixation map for all subjects. Some of these images have small salient objects, and others have large-scale regions of interest. Analyzing the qualitative results, we can see that our method shows more resemblance to the ground truth than the other 8 saliency approaches. The regions highlighted by our proposed method overlap to a surprisingly large extent with those image regions looked at by humans in free viewing. Good performance concerning color pop-out is also observed with our method as compared to other approaches. MSS, HC, and IT can obtain fine resolution saliency maps, but they are more likely to focus on large-scale structures and thereby miss some small salient objects.

Saliency Detection for Natural Images

In this subsection, we compare our method with 8 other saliency approaches on the Extended Complex Scene Saliency Dataset (ECSSD) dataset (Shi et al., 2016) that includes 1,000 natural images and corresponding GT images. **Figure 4** gives the saliency maps for eight sample images from the ECSSD dataset, which provide a visual comparison of all saliency methods. It can be seen that MSS, HC, and IT can obtain high-resolution saliency maps, but they suffer from cluttered backgrounds. PQFT, PCT, HFT, and BSWHT can detect small salient objects effectively, but sometimes they fail to highlight the whole salient objects with relatively large size. Note that our proposed method can enhance the salient regions and meanwhile suppress background clutters heavily. Moreover, since our method computes visual salience in a multiscale manner, it can detect both small and large scale salient regions simultaneously.

To evaluate the detection accuracy objectively, we plot the P-R curves and the ROC curves for all saliency approaches as shown in **Figures 5A,B**. Note that a high ROC or P-R curve indicates the saliency maps have a high resemblance with the GT images. As can be seen, our method and HFT obtain comparatively high curves as compared to other saliency approaches. Nevertheless, it can be noticed that our method is slightly better than HFT. **Table 2** lists the ROC-AUC score averaged over all 1,000 images for each saliency method. As expected, our method obtains the highest ROC-AUC score. This means that our method achieves the best performance in this saliency detection test.

It should be noted that this article mainly studies the computation of bottom-up visual saliency. Bottom-up attention or saliency studies mostly use psychophysical patterns (section Psychophysical Consistency) and Bruce and Tsotsos's eye fixation prediction dataset (section Eye Fixation Prediction). These two datasets were created specifically for the bottom-up attention

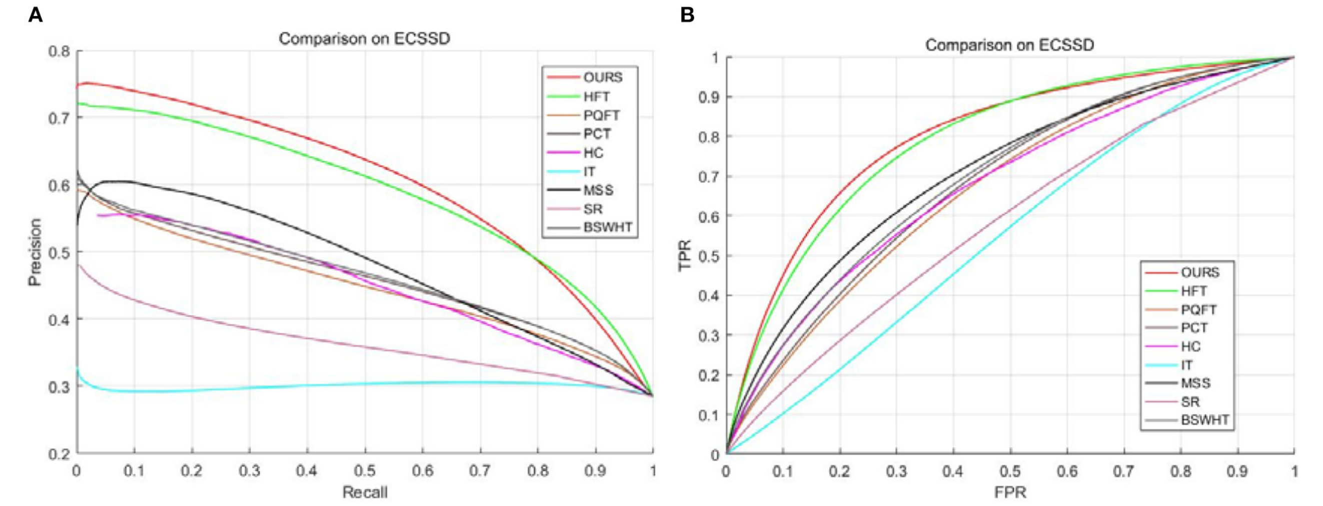


FIGURE 5 | (A) P-R curves on the ECSSD dataset. **(B)** ROC curves on the ECSSD dataset.

TABLE 2 | The ROC-AUC scores of all nine saliency methods.

Method	Ours	SR	PQFT	PCT	HFT	BSWHT	DN	MSS	HC	IT
AUC	0.7990	0.5805	0.6681	0.6813	0.7895	0.6954	0.6333	0.7091	0.6755	0.5493

tests. For more testing, we conducted tests on the ECSSD dataset in this section. The ECSSD dataset is particularly used for foreground region segmentation methods. They are not only purely bottom-up but also need more top-down calculations. Nevertheless, our method still achieves good performance in this test.

Computational Time Cost

Computational speed is an important metric to evaluate the performance of a saliency model. We also record the computational time cost per image from the ECSSD dataset in a standard desktop computer environment. **Table 3** gives each method's Matlab runtime measurements averaged over the data set. It can be seen that the traditional frequency-domain models (SR, PQFT, PCT, and BSWHT) are relatively faster than other methods. As a multiscale frequency domain calculation model of visual saliency, our method needs about 10 times the computational cost of the traditional frequency-domain model. Nevertheless, it has about the same computational cost as HFT and MSS and is still faster than HC and IT. Note that all saliency methods are implemented on such a computer platform as Intel i7-8650U 1.90GHz CPU, and 16GB of memory.

APPLICATIONS TO SHIP DETECTION IN OPTICAL SATELLITE IMAGES

In this section, we apply the proposed method to detect ship signatures in optical satellite images. To validate the effectiveness of our method, we conduct experiments by use of real optical satellite images from the MASATI dataset (Antonio-Javier et al., 2018). All tests in this section are run on a Windows platform (Microsoft Incorporation, US). The computer is equipped with a quad-core Intel 2.9 GHz CPU and 32 GB of memory (Intel Incorporation, US). All the program codes are implemented in the MATLAB (MathWorks Incorporation, US) R2017b environment.

Saliency-Based Ship Detection in Optical Satellite Images

Automatic ship detection in optical satellite images has attracted intensive investigations (Bi et al., 2012; Jubelin and Khenchaf, 2014; Qi et al., 2015; Zou and Shi, 2016; Li et al., 2020). It plays a crucial role in a maritime surveillance system. Some studies perform ship detection by using synthetic aperture radar (SAR) (Crisp, 2004; Yu et al., 2011a). However, strong speckles (caused by the coherence of backscattered signals) pose great difficulties for an automatic ship detection system. Compared with the SAR data, optical satellite images can provide more detailed characteristics of ship signatures.

More often than not, automatic ship detection will encounter two challenges. First, a marine surveillance system needs fast algorithms since it has to deal with a large amount of data in real-time. Second, lots of background disturbances always exist in the optical satellite images. Conventional target detectors use a constant false alarm rate (CFAR) which automatically adapts to the statistical distribution of sea clutters and targets of interest (Chen and Reed, 1987; Reed and Yu, 1990; Yu and Reed, 1993). However, if the signature of a target has similar intensities as its

TABLE 3 | Computational time cost per image for all saliency methods over the ECSSD dataset.

Method	Ours	SR	PQFT	PCT	HFT	BSWHT	MSS	HC	IT
Time(s)	0.1272	0.0109	0.0163	0.0147	0.1183	0.0102	0.0934	0.3349	0.2224

surroundings, the CFAR detector cannot discriminate the targets from their background clutters. It should be noticed that human vision is superior to existing techniques in observing a slick in the surrounding sea, and some vessels undetected by conventional algorithms are visible to the eye. Motivated by this fact, we employ our proposed saliency method to perform ship detection in optical satellite images.

Since ships are visually salient and will become dominant locations in a saliency map, a constant threshold value can be employed to discriminate the ship targets from sea backgrounds. However, a constant threshold will produce false alarms when no ship target appears in the scene under view. Therefore, we consider designing an adaptive threshold to detect the ship targets. The threshold value is computed by using the saliency values of the given saliency map:

$$T_s = \alpha(\mu_s + 2\sigma_s) \qquad (13)$$

where μ_s and σ_s are, respectively, the mean value and the standard deviation of the final saliency map, and α is an empirically tuned parameter. Note that a small α may lead to false alarms although it can detect ship targets; whereas a large α is likely to miss some ship signatures although it avoids false alarms. Through lots of experiments, we find that the detection results are reasonable when the parameter $\alpha = 4$.

An important note about our method's application to ship detection in optical satellite images is that the saliency map should be computed at full resolution. This is different from the salience computation for a natural image. Note that great disparities may exist in the size of various ships, and our method can detect both small and large salient objects simultaneously when the saliency map is computed at a high resolution. Therefore, to obtain high-resolution saliency maps with well-defined boundaries of targets, we directly use full-resolution optical satellite images to compute their saliency maps. This computation process can be considered as a human looking at the scenes at a fine resolution in a very careful manner.

Test on the MASATI Dataset

We conduct our method over the MASATI dataset that contains 6,212 satellite images in the visible spectrum. The dataset was collected from Microsoft Bing Maps (Antonio-Javier et al., 2018), of which each image has been manually labeled according to various classes. Since our tests only concern ship detection from sea backgrounds, we choose three sub-classes: *ship*, *multi,* and *detail* to test our multiscale saliency-based ship detection method. The *ship* sub-class represents images where a single ship appears within the image. The *multi* sub-class describes other images in which two or more instances of ships appear within them. In both sub-classes, the ships have lengths between 4 and 10 pixels. The *detail* sub-class are images with large-scale ships within a length between 20 and 100 pixels. The images were captured in RGB, and the average image size has a spatial resolution of around 512 × 512 pixels. The dataset was compiled between March and September of 2016 from different regions in Europe, Africa, Asia, the Mediterranean Sea, and the Atlantic and Pacific Oceans. We cannot provide simultaneous ground truths at present; nevertheless, the referred targets can be visually interpreted from these optical satellite images. Some typical test results are shown in **Figures 6**, **7**.

Figure 6 shows six sample images with a single ship target from the *ship* and the *detail* sub-classes of the MASATI dataset. The images contain disturbances of ship wakes, sea waves, clutters, and heterogeneities, which will cause challenges for a ship detection task. The 2nd−10th rows of **Figure 6** present the saliency maps of 7 comparison saliency approaches, the detection results of CFAR, and the saliency maps and detection results of our method, respectively. It can be seen that PQFT, PCT, BSWHT, HC, and IT cannot suppress the background disturbances effectively, particularly for the images with small ships. Although PQFT, BSWHT, MSS, HC, and IT can detect large ships, they fail to uniformly highlight the whole salient regions for these large-scale targets. It seems that HFT finds both small and large targets in this test, but it highlights some heterogeneous regions in the 3rd image. The CFAR method fails to detect small ship targets whereas it causes false alarms even though it works at a low false alarm rate. It should be noted that both small and large ship locations in our saliency maps can pop out relative to the clutter backgrounds and therefore are successfully detected by our method.

Figure 7 shows six sample images with multiple ship targets from the *multi* and the *detail* sub-classes of the MASATI dataset. This test is somewhat difficult because the sample images comprise strong disturbances including reefs, ship wakes, cloudlets, heterogeneities and clutters of seawater, etc. Moreover, there may exist a huge disparity in the size of the ships in a scene (5th and 6th images). The 2nd−10th rows of **Figure 7** present the saliency maps of 7 comparison approaches, the detection results of CFAR, the saliency maps, and the detection results of our method, respectively. Since PQFT, PCT, and BSWHT only compute visual saliency on a single scale, they cannot effectively suppress the background disturbances for these cluttered scenes. Note that HC, MSS, and IT compute visual salience in the spatial domain. They cannot suppress the cloudlets or other disturbances effectively. The CFAR detector inherently has numerous false alarms and cannot discriminate ships from these false alarms. It can be seen that our method highlights the ships and meanwhile suppresses the background disturbances in the saliency maps. Since our method can compute multiscale visual saliency, it accurately finds both small and large ship targets in this difficult test.

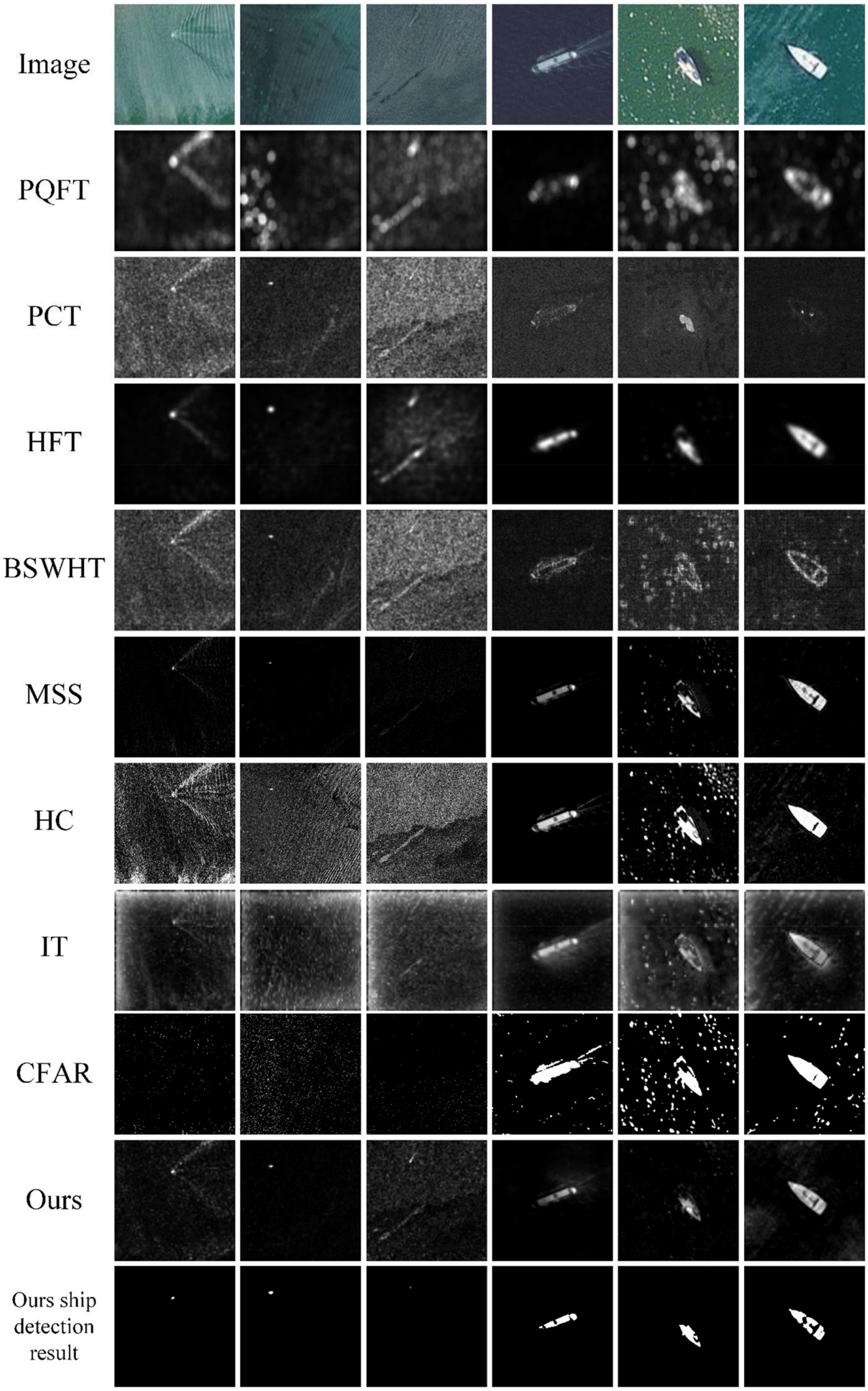


FIGURE 6 | A visual comparison of the saliency maps obtained by all eight approaches, as well as the detection results of CFAR and our method for the MASATI images with a single ship target.

CONCLUSION AND DISCUSSION

This article investigates automatic detection of bottom-up visual saliency from the perspective of multiscale analysis and computation in the frequency domain. We manifested that multiscale saliency information can be computed by performing multiscale wavelet decomposition and computation upon the magnitude coefficients in the frequency domain.

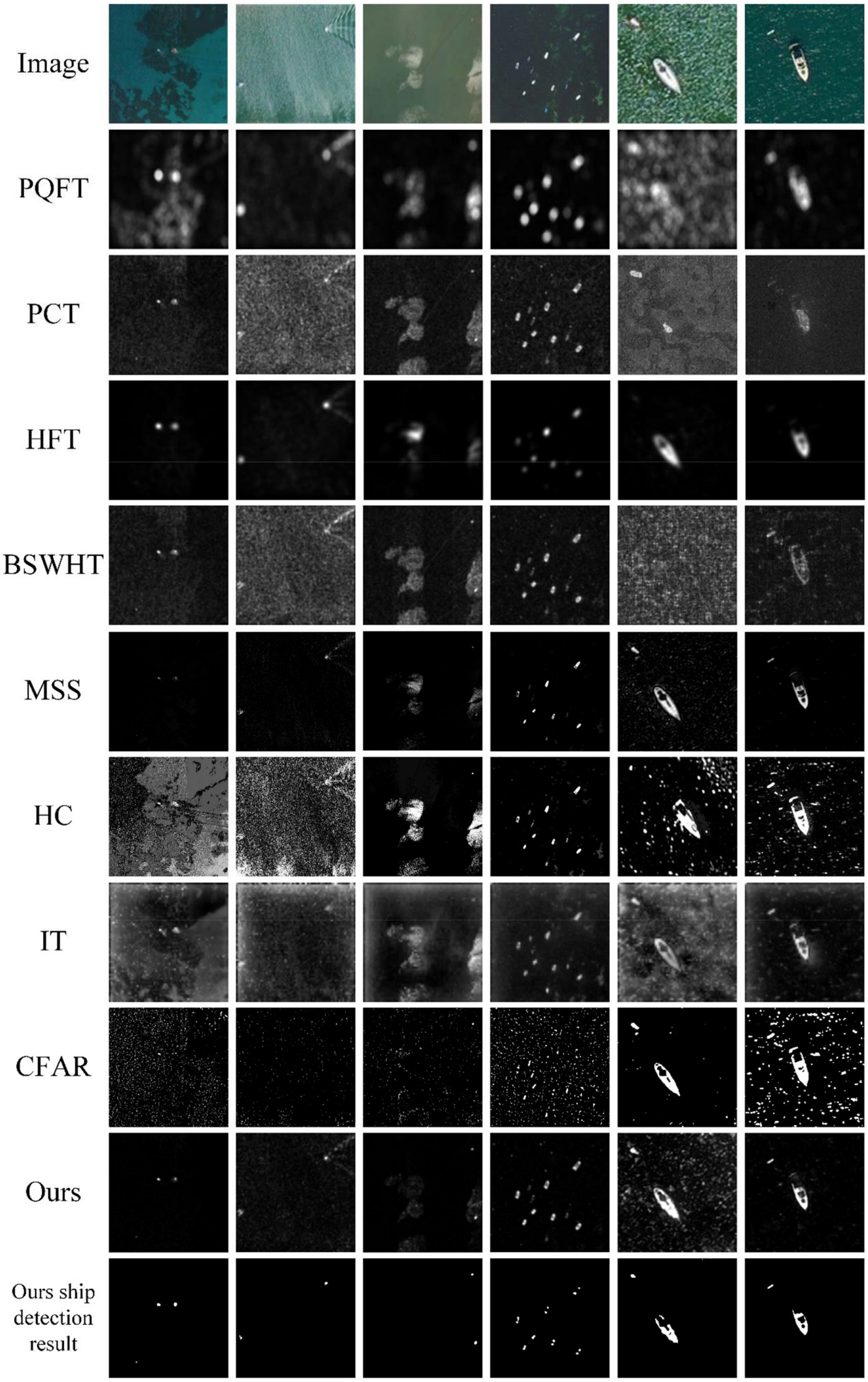


FIGURE 7 | A visual comparison of the saliency maps obtained by all eight approaches, as well as the detection results of CFAR and our method for the MASATI images with multiple ships.

The proposed model simulates the multiscale cortical center-surround suppression and has biological plausibility. The model is fast and can provide multiscale saliency maps, which are important for detecting salient objects of different sizes. Experiments over psychophysical patterns and natural image datasets showed that the proposed model outperforms state-of-the-art saliency approaches when evaluated by the ability to predict human fixations, and by the objective metrics of the

P-R curves and the ROC-AUC scores. The applications to ship detection in optical satellite images proved that the proposed multiscale visual saliency model is very effective in detecting both small and large ship targets simultaneously from the surrounding sea and robust against various background disturbances.

The main contribution of this article is to extend the traditional frequency-domain visual saliency model to multiscale saliency calculation. The traditional visual saliency model uses single-scale frequency domain calculation, while our new model uses multiscale frequency domain calculation. The multiscale visual saliency calculation is realized by decomposing the frequency domain coefficients of the input image by multiscale wavelet transform. The traditional frequency-domain calculation model has good detection ability for small targets, but weak detection ability for large targets. The advantage of our multiscale saliency calculation model is that it can calculate large-scale and small-scale saliency targets at the same time.

The limitation of this work is that it is only concerned with the detection of bottom-up visual saliency. It has not considered top-down influences such as some cues for selecting suitable scales of salience, or some cues for object recognition depending on a given vision task. Future work will focus on a task-dependent attention selection system. It is possible to add top-down influences for developing more intelligent vision systems to accomplish various visual search tasks in engineering applications.

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/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (62166048 and 61263048), and by the Applied Basic Research Project of Yunnan Province (2018FB102).

REFERENCES

Achanta, R., and Suesstrunk, S. (2010). "Saliency detection using maximum symmetric surround," in *International Conference on Image Processing* (Hong Kong), 2653–2656. doi: 10.1109/ICIP.2010.5652636

Antonio-Javier, G., Antonio, P., and Pablo, G. (2018). Automatic ship classification from optical aerial images with convolutional neural networks. *Remote Sens.* 10:511. doi: 10.3390/rs10040511

Bi, F., Zhu, B., Gao, L., and, Bian, M. (2012). A visual search inspired computational model for ship detection in optical satellite images. *IEEE Geosci. Remote Sens. Lett.* 9, 749–753. doi: 10.1109/LGRS.2011.2180695

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 185–206. doi: 10.1109/TPAMI.2012.89

Bruce, N., and Tsotsos, J. (2009). Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* 9, 1–24. doi: 10.1167/9.3.5

Chen, J. Y., and Reed, I. S. (1987). A detection algorithm for optical targets in clutter. *IEEE Trans. Aerosp. Electron. Syst.* 23, 46–59. doi: 10.1109/TAES.1987.313335

Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H. S., and Hu, S. M. (2015). "Global contrast based salient region detection," in *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 569–582. doi: 10.1109/TPAMI.2014.2345401

Crisp, D. J. (2004). *The State-of-the-Art in Ship Detection in Synthetic Aperture Radar Imagery*. Edinburgh, SA: Australian Government Department of Defence Defence Science and Technology Organisation, DSTO-RR-0272.

Davis, J., and Goadrich, M. (2006). "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, *Vol. 6* (Pittsburgh, PA: ACM), 233–240. doi: 10.1145/1143844.1143874

Guo, C., and Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Process.* 19, 185–198. doi: 10.1109/TIP.2009.2030969

Hou, X., and Zhang, L. (2007). "Saliency detection: a spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN), 1–8. doi: 10.1109/CVPR.2007.383267

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259. doi: 10.1109/34.730558

Jubelin, G., and Khenchaf, A. (2014). "Multiscale algorithm for ship detection in mid, high and very high resolution optical," in *2014 IEEE Geoscience and Remote Sensing Symposium* (Quebec City, QC: IGARSS), 2289–2292. doi: 10.1109/IGARSS.2014.6946927

Li, J., Levine, M. D., An, X., Xu, X., and He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 996–1010. doi: 10.1109/TPAMI.2012.147

Li, L., Zhou, Z., Wang, B., Miao, L., and Zong, H. (2020). A novel CNN-based method for accurate ship detection in HR optical remote sensing images via rotated bounding box. *IEEE Trans. Geosci. Remote Sens.* 59, 686–699. doi: 10.1109/TGRS.2020.2995477

Liu, G., and Yang, J. (2019). Exploiting color volume and color difference for salient region detection. *IEEE Trans. Image Process.* 28, 6–16. doi: 10.1109/TIP.2018.2847422

Qi, S., Ma, J., Lin, J., Li, Y., and Tian, J. (2015). Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images. *IEEE Geosci. Remote Sens. Lett.* 12, 1451–1455. doi: 10.1109/LGRS.2015.2408355

Reed, I. S., and Yu, X. (1990). Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Trans. Acoust. Speech Signal Process.* 38, 1760–1770. doi: 10.1109/29.60107

Shi, J., Yan, Q., Xu, L., and Jia, J. (2016). Hierarchical image saliency detection on extended CSSD. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 717–729. doi: 10.1109/TPAMI.2015.2465960

Tatler, B., Baddeley, R., and Gilchrist, I. (2005). Visual correlates of fixation selection: effects of scale and time. *Vis. Res.* 45, 643–659. doi: 10.1016/j.visres.2004.09.017

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Wolfe, J. M., and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501. doi: 10.1038/nrn1411

Wolfe, J. M., and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nat. Hum. Behav.* 1:0058. doi: 10.1038/s41562-017-0058

Yu, X., and Reed, I. S. (1993). Comparative performance analysis of adaptive multispectral detectors. *IEEE Trans. Signal Process*. 41, 2639–2656. doi: 10.1109/78.229895

Yu, Y., Wang, B., and Zhang, L. (2009). "Pulse discrete cosine transform for saliency-based visual attention," in *2009 IEEE 8th International Conference on Development and Learning* (Shanghai), 1–6.

Yu, Y., Wang, B., and Zhang, L. (2011a). Hebbian-based neural networks for bottom-up visual attention and its applications to ship detection in SAR images. *Neurocomputing* 74, 2008–2017. doi: 10.1016/j.neucom.2010.06.026

Yu, Y., Wang, B., and Zhang, L. (2011b). Bottom-up attention: pulsed PCA transform and pulsed cosine transform. *Cogn. Neurodyn.* 5, 321–332. doi: 10.1007/s11571-011-9155-z

Yu, Y., and Yang, J. (2017). Visual saliency using binary spectrum of Walsh–Hadamard transform and its applications to ship detection in multispectral imagery. *Neural Process. Lett.* 45, 759–776. doi: 10.1007/s11063-016-9507-0

Zhaoping, L. (2002). A saliency map in primary visual cortex. *Trends Cogn. Sci*. 6, 9–16. doi: 10.1016/S1364-6613(00)01817-9

Zhaoping, L., and Peter, D. (2006). Pre-attentive visual selection. *Neural Netw*. 19, 1437–1439. doi: 10.1016/j.neunet.2006.09.003

Zou, Z., and Shi, Z. (2016). Ship detection in spaceborne optical image with SVD networks. *IEEE Trans. Geosci. Remote Sens*. 54, 5832–5845. doi: 10.1109/TGRS.2016.2572736

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/fnbot.2021.785808

ShapeEditor: A StyleGAN Encoder for Stable and High Fidelity Face Swapping

*Shuai Yang, Kai Qiao, Ruoxi Qin, Pengfei Xie, Shuhao Shi, Ningning Liang, Linyuan Wang, Jian Chen, Guoen Hu and Bin Yan**

Henan Key Laboratory of Imaging and Intelligent Processing, People's Liberation Army (PLA) Strategy Support Force Information Engineering University, Zhengzhou, China

With the continuous development of deep-learning technology, ever more advanced face-swapping methods are being proposed. Recently, face-swapping methods based on generative adversarial networks (GANs) have realized many-to-many face exchanges with few samples, which advances the development of this field. However, the images generated by previous GAN-based methods often show instability. The fundamental reason is that the GAN in these frameworks is difficult to converge to the distribution of face space in training completely. To solve this problem, we propose a novel face-swapping method based on pretrained StyleGAN generator with a stronger ability of high-quality face image generation. The critical issue is how to control StyleGAN to generate swapped images accurately. We design the control strategy of the generator based on the idea of encoding and decoding and propose an encoder called ShapeEditor to complete this task. ShapeEditor is a two-step encoder used to generate a set of coding vectors that integrate the identity and attribute of the input faces. In the first step, we extract the identity vector of the source image and the attribute vector of the target image; in the second step, we map the concatenation of the identity vector and attribute vector onto the potential internal space of StyleGAN. Extensive experiments on the test dataset show that the results of the proposed method are not only superior in clarity and authenticity than other state-of-the-art methods but also sufficiently integrate identity and attribute.

Keywords: face swapping, generative adversarial network, disentanglement, style transfer, deepfake

OPEN ACCESS

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Durai Raj Vincent P.M,
VIT University, India
Yanan Guo,
Beijing Information Science and Technology University, China

****Correspondence:***
Bin Yan
ybspace@hotmail.com

Received: *29 September 2021*
Accepted: *23 November 2021*
Published: *21 January 2022*

Citation:
Yang S, Qiao K, Qin R, Xie P, Shi S, Liang N, Wang L, Chen J, Hu G and Yan B (2022) ShapeEditor: A StyleGAN Encoder for Stable and High Fidelity Face Swapping. Front. Neurorobot. 15:785808. doi: 10.3389/fnbot.2021.785808

1. INTRODUCTION

As one of the main contents of deepfake, face swapping declares to the world today that seeing is not always believing. Face swapping refers to transferring the identity of a source image to the face of another target image while keeping unchanged the illumination, head posture, expression, dress, background, and other attribute information of the target image. Face swapping has received widespread attention since its birth, catering to the affluent needs of social life, such as hairstyle simulation, film and television shooting, privacy protection, and so on (Ross and Othman, 2010).

Face swapping is accompanied not only by its interesting and operational application prospects but also by various challenges between reality and vision. The early face-swapping methods (Bitouk et al., 2008; Korshunova et al., 2017) require many images of source and target characters to provide sufficient facial information. Otherwise, the models would not have a suitable reference basis to produce good results. Some three-dimensional-based (3D-based) methods (Olszewski et al., 2017; Nirkin et al., 2018; Sun et al., 2018) make use of the advantage of fitting 3D face models to deal with the problems of large angle and small samples. At the same time, due to the limited accuracy of 3D face models, it is impossible to generate works with better details and higher fidelity. Recently, with the continuous tapping of the potential of generative adversarial networks (GANs) (Nandhini Abirami et al., 2021), some face-swapping methods based on GANs (Bao et al., 2018; Natsume et al., 2018a,b; Li et al., 2019; Nirkin et al., 2019) can achieve a good fusion of identity and attribute information with only a small number of samples, reflecting the effect of great creativity. Unfortunately, the surprising creativity of these methods does not offset the adverse impacts of their frequent artifacts and low-resolution limitation.

On another track, the most advanced face image generation methods have generated facial images with high resolution and realistic texture. Most notably, StyleGAN (Karras et al., 2019) can randomly generate a variety of clear faces with a resolution of up to 1024 × 1024. StyleGAN has three potential spaces: initial potential space $\mathcal{Z}$, intermediate potential space $\mathcal{W}$, and extended potential space $\mathcal{W}+$. (Abdal et al., 2019) proved that the concatenation of 18 different 512-dimensional vectors is the easiest way to embed an image and obtain a reasonable result. On this basis, various works (Gu et al., 2020; Härkönen et al., 2020; Richardson et al., 2020; Zhu et al., 2020) explore in detail the StyleGAN potential vector space: some (Shen and Zhou, 2020; Shen et al., 2020; Tewari et al., 2020) find a linear direction to control the change of a single facial attribute, some (Nitzan et al., 2020) control facial expression and posture in the original StyleGAN image domain, and others (Richardson et al., 2020; Wang et al., 2021) deal well with the difficult task of facial super-resolution.

In contrast with other face-swapping methods, the first criterion we pursue is that the images after face swapping have both higher clarity and better authenticity. We propose a many-to-many face-swapping method based on the pretrained StyleGAN model (Karras et al., 2019), which strives to ensure the clarity and fidelity of the results while fusing identity and attribute information. Given the inherent ability of the pretrained StyleGAN model to generate random high-quality face images, the difficulty of this task is how to accurately render the corresponding latent vectors. To achieve this goal, we first designed an encoder, ShapeEditor, to find the corresponding codes in the $\mathcal{W}+$ vector space. The workflow of the encoder was divided into two stages, the first being the respective extraction of identity and attribute codes, and the second being to map the combination of two-channel codes into the potential input vector domain of the pretrained model. Moreover, we designed a set of loss functions with a strong monitoring ability to urge ShapeEditor to update parameters to learn to map step by step onto the latent space of StyleGAN. As verification, we made numerous qualitative and quantitative experimental comparisons with the existing face-swapping methods, which show the unique advantages of the proposed method.

2. RELATED WORKS

Recently, the GAN-based face-swapping methods have shown better performance, thus attracting more extensive research and attention. Although integrate attributes and identity information well, these methods generally have the common problem of poor clarity and authenticity. On the other hand, as GAN with better image quality has been proposed, many works are devoted to manipulating GAN's semantic space to generate clear and stable images. We creatively combine the advantages of the above two fields to improve the performance of face swapping, and make possible the more complex control of GAN's potential space.

2.1. GAN-Based Face Swapping

Olszewski et al. (2017) fit the 3D face model of the source face and used a conditional generator of the coder-decoder structure to infer the converted face texture. Too simple generator network structure and training strategy make this method unable to separate identity and attribute information to further complete many-to-many identity exchange. Sun et al. (2018) trained a convolutional neural network to regress the parameters of a 3D model of the input face, replaced the identity parameters, and combined the region around the head to generate a realistic face-swapped image. Limited to the accuracy of the model reconstruction, 3D-based face-swapping methods are unsatisfactory in terms of attribute and identity fidelity. Face Swapping GAN (FSGAN) (Nirkin et al., 2019) used sparse landmarks to track facial expression, and designed GANs with different functions for the three stages of face swapping. This method realized subject agnostic face swapping, while being limited by the resolution of the input image and the complexity of expression. Bao et al. (2018) implemented this task using a more concise coder-decoder architecture, in which two independent coders separate the identity and attributes of human faces. This method used an asymmetric training strategy to promote a large number of unlabeled faces to contribute to the training. Following the basic network framework and asymmetric training strategy of Bao et al. (2018), FaceShifter (Li et al., 2019) has done meaningful work on embedding multi-level information in the generator and handling occlusion more robustly. The generator leverages denormalizations for feature integration in multiple feature levels, showing a better representation of identity and attribute. However, the clarity and stability of the image generated by FaceShifter are not always ideal. As shown in **Figure 1**, the eyebrows of the result in the first line appear ghosting, and the nose of the result in the second line appear artifact. These examples show that the most advanced GAN-based face-swapping method is still insufficient in authenticity.

2.2. The Potential and Challenge of Pretrained GAN Manipulation

While a lot of works have been done on how to control GAN to perform complex image operations, such as face swapping, others focus on improving the quality of images. Through carefully designed style-based network structure and layer-by-layer training, StyleGAN (Karras et al., 2019) realized high-definition and high-quality face image generation. With the help

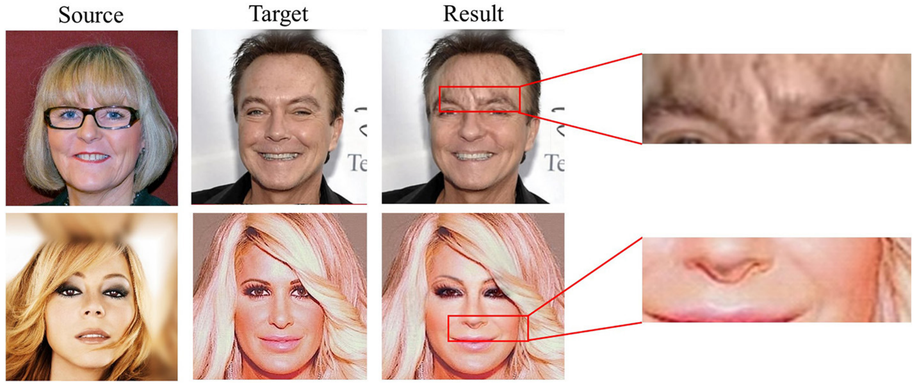


FIGURE 1 | Some abnormal results generated by FaceShifter (Li et al., 2019).

of pretrained StyleGAN, image quality is easier to be improved. The manipulation of StyleGAN is a difficult task, and most early works are limited to understanding and reproducing the potential space of GAN. The inversion task of StyleGAN is to find the potential vector that best matches the given image. Abdal et al. (2019) took several minutes to embed a face into the StyleGAN image domain. Richardson et al. (2020), Zhu et al. (2020), and Gu et al. (2020) tried to improve efficiency using encoder structure, but the inversion results of wild images in their methods are unsatisfactory. Later, some more complex works appeared, such as changing individual attributes (smile, age, facial angle, etc.) (Härkönen et al., 2020; Shen and Zhou, 2020; Shen et al., 2020), establishing relationship between 3D semantic parameters and genuine facial expressions (Tewari et al., 2020), and super-resolution of low-quality facial images (Wang et al., 2021). To the best of our knowledge, there is no face-swapping method based on StyleGAN. This task requires more complex semantic manipulation, and the current controllers are not competent. Nitzan et al. (2020) did closely related work to control expression through latent space mapping. However, working in the $\mathcal{W}$ space led to the failure of embedding wild images into potential space. In addition, the single vector of the attribute is too plain to carry the information of background, posture, expression, etc.

2.3. The Inheritance and Transcendence

We propose a StyleGAN encoder, called ShapeEditor, for stable and high-fidelity face swapping. As the combination of face swapping and pretrained GAN manipulation, ShapeEditor inherits and surpasses the latest ideas in the two fields.

We use an asymmetric training strategy similar to that in FaceShifter (Li et al., 2019) to realize the training process without labeled data, so as to ensure solid constraints and reduce data processing costs. Moreover, the well-designed coder-decoder structure of our framework can firmly guarantee image quality, which is the weakest aspect of FaceShifter. Inspired by SPADE (Park et al., 2019) and AdaIN (Huang and Belongie, 2017), the FaceShifter generator designs AAD layer-level denormalization for feature integration in multiple feature levels. By comparison, the internal mapper of ShapeEditor is composed of lightweight Multilayer Perceptrons (MLP) to generate feature vectors embedded in StyleGAN $\mathcal{W}+$ space, which reduces the burden of model training.

Our method and Nitzan et al. (2020) both use the decoupling framework to extract attribute and identity code through attribute extractor and identity extractor, respectively. The codes are then mapped into the latent space of the employed pretrained generator. Our key difference is that we select $\mathcal{W}+$ potential space as the mapping space, which is the premise of realizing the complex semantic operation of face swapping. In addition, in order to recover the attribute information more finely, we use multi-level feature mapping instead of a single output as attribute code like Nitzan et al. (2020) did. The ablation study proves that our pertinent designs make a significant contribution to better semantic manipulation.

3. METHODS

Our method requires two images as input: I_{attr} and I_{id}. We expect the output of the model to reflect the identity of I_{id} and the facial expression, head posture, hairstyle, lighting, and other attribute information of I_{attr}. Therefore, the main challenge of this work is to obtain the StyleGAN potential vectors that are consistent with the $\mathcal{W}+$ spatial distribution and better integrate attributes and identity. To solve this problem, we designed a two-step coding process. As shown in **Figure 2A**, the entire mapping process is divided into two phases: ID-ATTR encoding and latent-space encoding. In the first stage, E_{id} extracts the identity vector of I_{id}, and E_{attr} extracts the attribute vector of I_{attr}. As shown in **Figure 2B**, inspired by pSp (Richardson et al., 2020), E_{attr} consists of a pyramid-shaped three-layer feature map extraction structure and a set of convolutional mappers (CM). In the second stage, we input the concatenation of $E_{\text{id}}(I_{\text{id}})$ and $E_{\text{attr}}(I_{\text{attr}})$ into the multilayer perceptron (MLP) of each layer and map the vectors containing identity and attribute information directly to the $\mathcal{W}+$ potential vector space. In summary, the whole image conversion process can be represented as

$$I_{\text{out}} = G\Big(MLP\big([E_{\text{id}}(I_{\text{id}}), E_{\text{attr}}(I_{\text{attr}})]\big)\Big), \qquad (1)$$

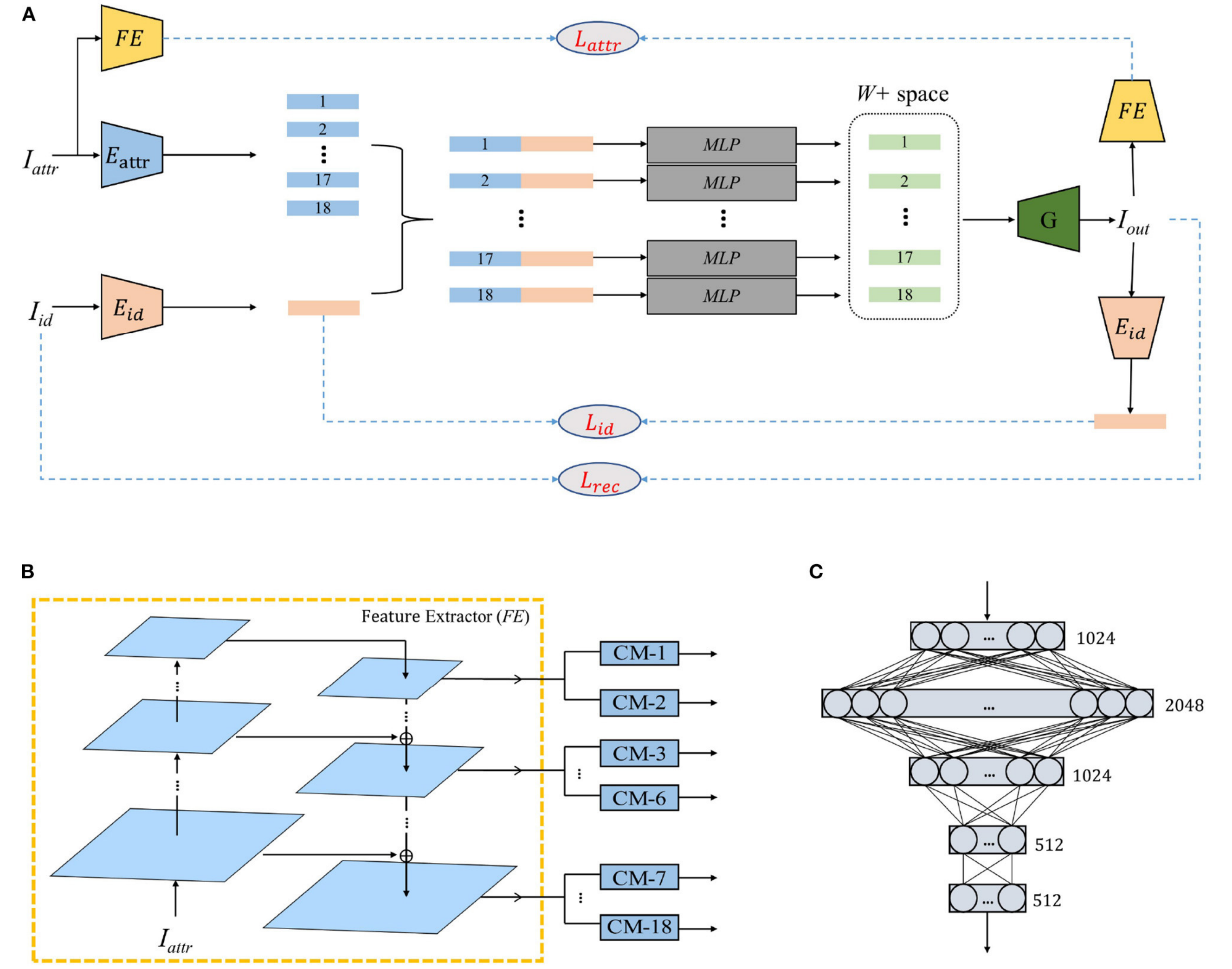


FIGURE 2 | The overall structure and data flow of the proposed model. **(A)** is the flow of our method. **(B)** is the structure of E_{attr}. **(C)** is the structure of Multilayer Perceptron (MLP).

where $G(\cdot)$ is the pretrained StyleGAN model, $MLP(\cdot)$ is the multilayer perceptron, and $[\cdot,\cdot]$ is the concatenation of two vectors.

3.1. Network Architecture

E_{id} is pretrained ArcFace (Deng et al., 2019) model. We use ResNet-IR (Deng et al., 2019) for Feature Extractor (*FE*), in which the feature output layers are 27, 30, and 44. The CM is a fully convolutional network that compresses the tensor of 8 × 8 × 512 dimensions into 1 × 1 × 512 dimensions through three convolution operations with a step size of two. As shown in **Figure 2C**, *MLP* is a five-layer fully connected network. The StyleGAN generator is a pretrained model trained on FlickrFaces-HQ (FFHQ) (Karras et al., 2019).

We mainly use convolution to reduce the dimensions of image encoding and use deconvolution to decode $\mathcal{W}+$ vectors. E_{attr} and E_{id} achieve the data-dimension reduction from image to vector through convolution and other network operations. The identity vector and attribute vector dimensions are both 1 × 512. The splicing of identity and attribute vectors is then input into a set of MLP to convert the face style and map the low-dimensional information to $\mathcal{W}+$ space. The deconvolution process is mainly reflected in StyleGAN, which changes from vectors in $\mathcal{W}+$ space to images. Note that we do not change any structure of StyleGAN but hope to use its powerful image-generation capabilities to make our face-changing images more stable and clear.

3.2. Training and Loss Functions

The advanced face-recognition model accurately identifies the face, so we believe that it can extract face-feature information and take the feature vector extracted by the pretrained ArcFace (Deng et al., 2019) as the identity information. To ensure that the

```
Algorithm 1 Training ShapeEditor using gradient descent.
Input:
    $I_{attr}$ : Image containing attribute information
    $I_{id}$ : Image containing identity information
    $P$ : Identity-attribute image pair space
Functions:
    Encoder ShapeEditor: $P \rightarrow \mathcal{W}+$
    Generator G: $\mathcal{W}+ \rightarrow I$
    Loss ← $\mathcal{L}_{id}$ : Calculate the identity loss between $I_{id}$ and $I_{out}$.
    Loss ← $\mathcal{L}_{attr}$ : Calculate the attribute loss between $I_{attr}$ and $I_{out}$.
    Loss ← $\mathcal{L}_{rec}$ : Calculate the reconstruction loss between $I_{id}(I_{attr})$ and $I_{out}$.
Output:
    $I$ : Image space
    $\mathcal{W}+$ : Potential vector space of StyleGAN
    $I_{out}$ : Synthesized face-swapping image

for number of training iterations do:
    for $I_{id}$, $I_{attr}$ randomly selected in training dataset do:
        Generate the $\mathcal{W}+$ space vector using $[I_{id}, I_{attr}]$
                ShapeEditor: $P \rightarrow \mathcal{W}+$
        Generate the face-swapping image $I_{out}$ using the $\mathcal{W}+$ space vector
                G: $\mathcal{W}+ \rightarrow I$
        Calculate the identity loss $\mathcal{L}_{id}$, the attribute loss $\mathcal{L}_{attr}$, and the reconstruction loss $\mathcal{L}_{rec}$
        Update ShapeEditor with loss
    end
end.
```

identity of I_{out} is consistent with I_{id}, we introduce the identity loss

$$\mathcal{L}_{id} = \|E_{id}(I_{id}) - E_{id}(I_{out})\|_2, \tag{2}$$

where $E_{id}(\cdot)$ is the pretrained ArcFace model.

Similarly, we adopt certain restrictions to ensure that the attribute information of I_{out} is consistent with that of I_{attr}. Given that the three-layer feature map extraction structure should gradually have the ability to extract attribute information with the training process, we define the attribute loss function as

$$\mathcal{L}_{attr} = \|P(I_{attr}) - P(I_{out})\|_2^2, \tag{3}$$

where $P(\cdot)$ is the extraction structure.

Note that the attribute information of I_{attr} and the identity information of I_{id} should not only exist in I_{out} but should also be well integrated. Based on this idea, we define the reconstruction loss as

$$\mathcal{L}_{rec} = \begin{cases} \|I_{out} - I_{id}\|_2 + \|F(I_{out}) - F(I_{id})\|_2 & \text{if } I_{id} = I_{attr} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $F(\cdot)$ is the perceptual feature extractor in the loss of learned perceptual image patch similarity (Zhang et al., 2018), which extracts the perceptual information of the image at the high-dimensional level. $\mathcal{L}_2$ loss measures the difference between the two images at the pixel level. Note that $\mathcal{L}_{rec}$ has a positive value only when I_{id} and I_{attr} are the same because only in this case should I_{out} and I_{id} (or I_{attr}) be so consistent that they are exactly the same; otherwise, we cannot expect a similar comparison between the two images. Overall, our total training loss is the weighted sum of all the losses mentioned above:

$$\mathcal{L}_{total} = \lambda_{id}\mathcal{L}_{id} + \lambda_{attr}\mathcal{L}_{attr} + \lambda_{rec}\mathcal{L}_{rec}. \tag{5}$$

Based on the loss functions and model structure proposed above, we train the ShapeEditor encoder according to **Algorithm 1**.

4. EXPERIMENTS

Implementation Details: We use the FFHQ (Karras et al., 2019) dataset as the training set, and the value of loss weights is set to $\lambda_{id} = 0.5, \lambda_{attr} = 0.1, \lambda_{rec} = 1$. The ratio of the training data with $I_{id} = I_{attr}$ to that with $I_{id} \neq I_{attr}$ is set to $2:1$. During the training, the network parameters of E_{id} and the StyleGAN generator remain unchanged, and the weights of the rest are updated with iterations. To compare with other methods, we train the model with images of 256 × 256 resolution in this section. This model was trained on a single NVIDIA TITAN RTX for about 2 days with a Ranger optimizer (Richardson et al., 2020), with a batch size set to eight and a learning rate set to 0.0001.

4.1. Qualitative Comparison With Previous Methods

We compare the proposed method with FSGAN (Nirkin et al., 2019), FaceShifter (Li et al., 2019; Nitzan et al., 2020) on the CelebAMask-HQ (Lee et al., 2020) test dataset. **Figure 3** shows, as expected because the proposed method is based on a pretrained StyleGAN (Karras et al., 2019) with high-quality face-generation capabilities, that all the generation results (**Figure 3**, column 6) are stable and clear enough that there are no errors such as artifacts and abnormal illumination.

Almost every output image (**Figure 3**, column 3) of FSGAN (Nirkin et al., 2019) shows unnatural lighting transition and lack of facial details, the abnormal region of the face is caused by directly extracting and filling the internal area of the face (**Figure 3**, row 3, column 4), which is completely avoided in the proposed method.

Because there is no pretrained model as the backbone, it is difficult for FaceShifter (Li et al., 2019) to avoid facial blur, some results even show facial illumination confusion (**Figure 3**, row 3, column 4) and eye ghosting (**Figure 3**, row 7, column 4), showing that its authenticity is significantly inferior to that of the proposed method.

Similar to the proposed method, Nitzan et al. (2020) use StyleGAN (Karras et al., 2019) as the backbone. However, it cannot accurately integrate identity and attribute information because of its simple encoder structure and the constraint of $\mathcal{W}$ potential space. Therefore, although it can generate high-quality images (**Figure 3**, column 6), it is not as good as the proposed method for fusing semantic information, which is reflected in the

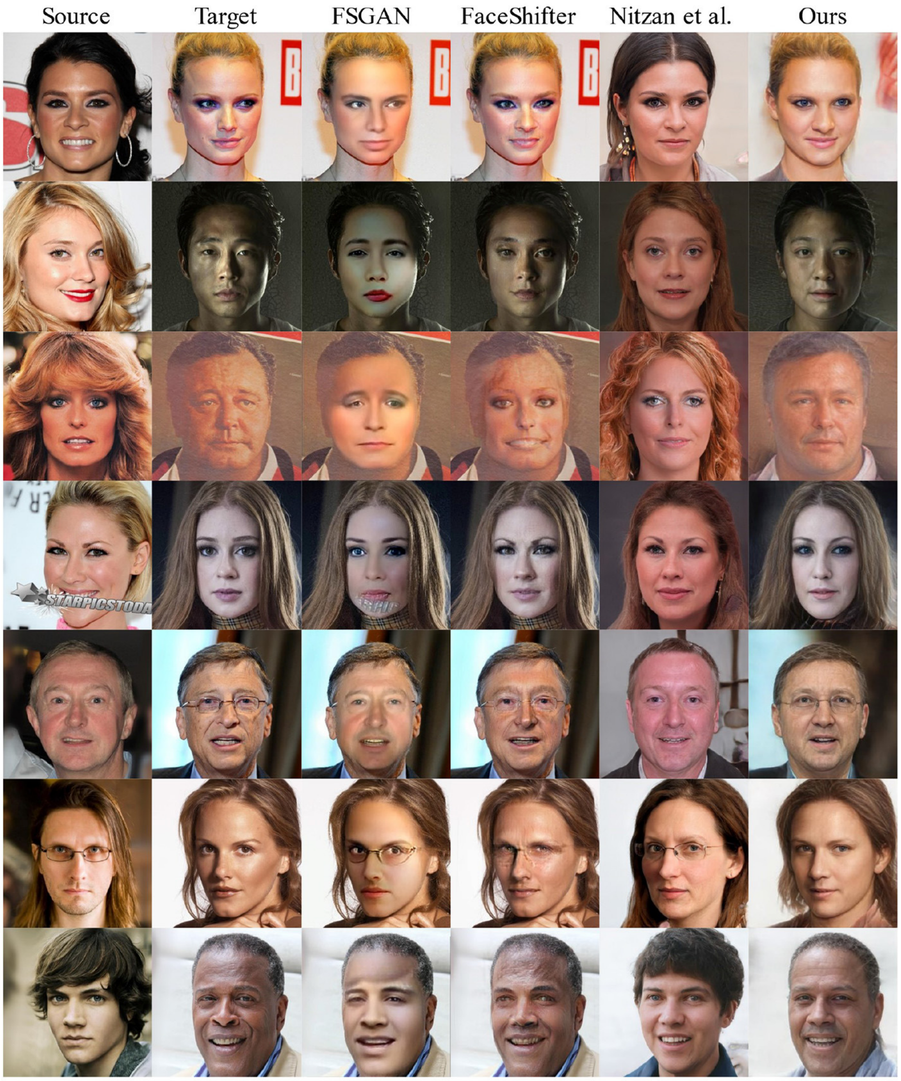


FIGURE 3 | Qualitative comparison with FSGAN (Nirkin et al., 2019), FaceShifter (Li et al., 2019; Nitzan et al., 2020) on the CelebAMask-HQ (Lee et al., 2020) test dataset.

attributes of the target image, such as hairstyle and background, that are not contained.

In addition to the excellent performance in terms of authenticity and fidelity, the proposed method also deals with extreme lighting conditions (**Figure 3**, row 2, column 6) and even keeps the sense of age (**Figure 3**, row 3, column 6). Thanks to that, we use the facial recognition module to extract the identity vector instead of directly using the pixels in the facial area. We can extract the identity information very well even if the source image has facial occlusion (**Figure 3**, row 4, column 6). The proposed model understands whether its output should have glasses (**Figure 3**, column 6, rows 5 and 6), which is embedded in the potential space of the pretrained StyleGAN model (Karras et al., 2019).

4.2. Quantitative Comparison With Previous Methods

As mentioned in the section 2.3, our method mainly inherits the ideas of latent space manipulation of pretrained models and GAN-based face swapping. To show the advantages, we compare the proposed method with other related. In the field of latent space manipulation, Nitzan et al. (2020) is the most similar to our work, which is about controlling facial attributes with StyleGAN. In the field of GAN-based face swapping, DeepFakes (Rössler et al., 2019), FSGAN (Nirkin et al., 2019), and FaceShifter (Li et al., 2019) occupy earlier positions and have achieved remarkable face exchange. To show the robustness of our method, we compare the proposed method with them quantitatively.

TABLE 1 | Quantitative comparison with Nitzan et al. (2020). Our method performs better in most indicators.

Method	Identity Error ↓		Pose Error ↓		Expression Error ↓		Mood Consistency ↑
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Acc. (%)
Nitzan et al. (2020)	**0.97**	0.30	5.99	7.16	10.13	5.38	65.35
Ours	1.30	0.33	**3.82**	6.88	**5.93**	3.63	**75.38**

Bold values represent the best. ↑ represents that the larger the value, the better. ↓ represents that the smaller the value, the better.

4.2.1. Comparison With Nitzan et al.

Our method and Nitzan et al. (2020) both make use of the image generation ability of pretrained StyleGAN, and make efforts to achieve adequate control of the human face. But we are different in the choice of mapping space and framework design. To show the significance of our improvement in semantic control, we quantitatively compare our method with Nitzan et al. (2020) in terms of identity, pose, expression, and mood consistency on CelebAMask-HQ (Lee et al., 2020) dataset.

The face swapping model not only needs to ensure the image quality but also needs to fuse the identity and attribute information to the greatest extent. We propose four indicators to measure these aspects. To calculate the identity information in the test stage, we use another advanced method called CurricularFace (Huang et al., 2020) as the face-recognition module to extract the identity vectors of source faces and face-swapping results, then use L2 distance to calculate the difference between them to get the identity error. To ensure that the conversion results are consistent with the target image in attribute, we use 3DDFA-V2 (Guo et al., 2020) to estimate the key face points and the head angle. For normalization, we use the two-dimensional (2D) coordinate information instead of 3D coordinate information to reduce the error impact of key-point estimation as much as possible, and calculate the average position of key points in each image, and then obtain the relative position of each point so as to establish a unified expression coordinate system. Based on the above, we take the difference between the target image and the resulting image in angle as pose error, in key face points as expression error. In addition to pose and expression, mood embodies the high-level semantics of face attribute. Inspired by Abirami and Vincent (2021), we use the emotion recognition model (Zhao et al., 2021) to detect the ability of face-swapping methods to transmit emotional information. Specifically, we recognize the moods of the swapped images and calculate the consistency of the mood recognition results before and after face exchange.

We randomly extract images from the CelebAMask-HQ dataset as source faces and take the remaining images as target faces to form one-to-one corresponding face combinations as the test dataset. As shown in **Table 1**, our method is superior to Nitzan et al. (2020) in pose error, expression error, and mood consistency, which shows our advantages in attribute information transfer. Our identity error is slightly higher than Nitzan, that is because face swapping brings more changes in head area than expression manipulation. Our advantages in most indicators demonstrate that we have realized better work in latent space manipulation.

TABLE 2 | Quantitative assessment with DeepFakes (Rössler et al., 2019), FSGAN (Nirkin et al., 2019), and FaceShifter (Li et al., 2019).

		DeepFakes	FSGAN	FaceShifter	Ours
Identity Error ↓	Avg.	1.35	1.51	**0.96**	1.30
	Std.	0.32	0.45	0.31	0.33
Pose Error ↓	Avg.	3.79	**2.81**	3.04	3.82
	Std.	1.99	4.41	6.70	6.88
Expression Error ↓	Avg.	8.82	5.03	**4.53**	5.93
	Std.	3.30	2.17	2.83	3.63
Mood Consistency ↑	Acc. (%)	39.80	72.77	**77.94**	75.38
SSIM ↓	Avg.	0.81	0.95	0.96	**0.75**
	Std.	0.09	0.03	0.03	0.08
PSNR ↓	Avg.	20.54	23.76	28.17	**20.22**
	Std.	2.60	2.30	1.92	1.62
FDR ↓	Tsd.=0.01	91.42	76.59	37.67	**15.18**
	Tsd.=0.05	83.83	48.99	11.66	**2.67**
	Tsd.=0.1	77.45	35.86	6.05	**1.09**
	Tsd.=0.2	70.86	24.22	2.86	**0.32**

Tsd. represents the threshold, which is set to judge whether samples are forged or not. Bold values represent the best. ↑ represents that the larger the value, the better. ↓ represents that the smaller the value, the better.

4.2.2. Comparison With Face Swapping Methods

To comprehensively show the face-swapping ability of our method, we conduct quantitative comparisons in transformation consistency and image quality with DeepFakes, FSGAN, and FaceShifter. Our work, FSGAN, and FaceShifter rely on a single reference or few references and are many-to-many approaches. At the same time, DeepFakes have to be supported by multi-images or videos to transfer faces in to two specific identities. Therefore, in order to ensure the effectiveness and efficiency of comparison, we extract DeepFakes conversion results from Rössler et al. (2019) dataset. The calculations of identity error, pose error, expression error, and mood consistency is the same as in section 4.2.1, which represent transformation consistency evaluation. Following the work of Yao et al. (2020), we employ peak signal-to-noise ratio (PSNR) (Huynh-Thu and Ghanbari, 2008) and structural similarity index (SSIM) (Wang et al., 2004) to measure the image reconstruction similarity between the target face and swapped face. Last but not least, to evaluate the clarity and authenticity of images, we use Li and Lyu (2018), which can effectively capture the artifacts in the forged images, to identify fake faces according to the resolution of the generated images. Specifically, we calculate the Forgery Detection Rate (FDR) of the output images. In the analysis of section 4.1, we know

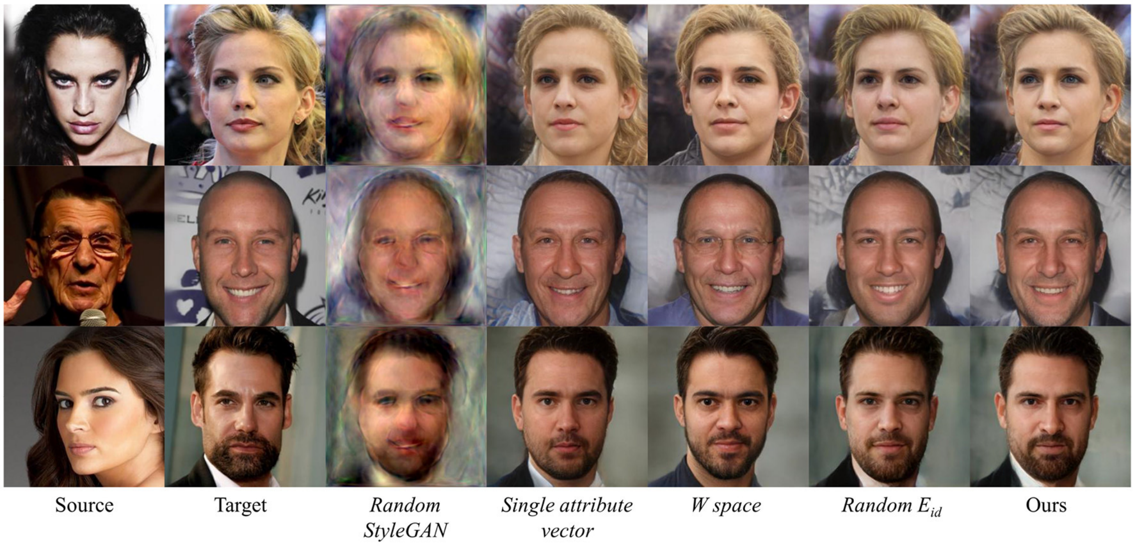


FIGURE 4 | Qualitative ablation study on different variants. Our original model performs better than others.

that the problems of low-quality images are mainly reflected in insufficient resolution and abnormal artifact areas. Therefore, the method of Li and Lyu (2018) can evaluate the quality of face images to a certain extent.

Table 2 lists the comparison results of different face-swapping methods. Notably, our method performs best in SSIM, indicating that our method retains the brightness, contrast, and structure of the original images to the greatest extent. Besides, our method outperforms others in PSNR, which demonstrates that our method can better preserve the global similarity than others. Also, our method has the least scores in FDR under different thresholds, which implies that our method can generate images with more sufficient resolution and less abnormal artifact areas. Finally, it is worth noting that our method has the second-best or the same level scores in identity error, pose error, expression error, and mood consistency, indicating that our method is comparable to others in identity and attribute, while being superior to them in terms of image quality and stability.

4.3. Ablation Study

To verify the effectiveness of each component of the proposed method, we do the ablation study by evaluating the following degenerate models of our method:

- *Random StyleGAN*. Using randomly initialized StyleGAN instead of pretrained generator.
- *Single attribute vector*. This variant uses a single output layer of Feature Extractor (*FE*), while the original uses multi-layer attribute information.
- *$\mathcal{W}$ space*. Using $\mathcal{W}$ potential space instead of $\mathcal{W}+$.
- *Random E_{id}*. Using randomly initialized E_{id} instead of pretrained face recognition model, with weight updating.

We report the qualitative results of the variants of our method in **Figure 4**. We can see that our original model has better face-swapping results. The results of *Random StyleGAN* are too vague to recognize, indicating that the pretrained StyleGAN can help to generate clear and vivid faces. The results of *Single attribute vector* lose details of hair, wrinkles, and beard compared with ours, showing that multi-layer *FE* can deliver more attribute information. The results of *$\mathcal{W}$ space* leak identity information and add unnecessary details like glasses, showing that $\mathcal{W}+$ potential space can more strictly embed wild faces into StyleGAN semantic space. The results of *Random E_{id}* leak identity information, which implies that using pretrained identity recognition model is of great significance.

TABLE 3 | Quantitative ablation study on different variants for face swapping.

		Single attribute vector	*$\mathcal{W}$ space*	*Random E_{id}*	Ours
Identity Error ↓	Avg.	**1.29**	1.33	1.37	**1.29**
	Std.	0.32	0.33	0.34	0.33
Pose Error ↓	Avg.	3.94	4.53	4.05	**3.64**
	Std.	5.59	5.74	5.75	5.55
Expression Error ↓	Avg.	6.63	7.43	6.48	**5.96**
	Std.	3.45	4.20	3.87	3.22
PSNR ↓	Avg.	19.38	**18.42**	19.94	20.22
	Std.	1.51	1.58	1.59	1.62
SSIM ↓	Avg.	0.73	**0.70**	0.74	0.75
	Std.	0.08	0.09	0.07	0.08

Bold values represent the best. ↓ represents that the smaller the value, the better.

Table 3 shows the quantitative results of the variants of our method on the randomly selected data from Lee et al. (2020) dataset. With the help of $\mathcal{W}+$ space and pretrained E_{id}, ours and *Single attribute vector* obtain lower identity error. The results of *$\mathcal{W}$ space* are much inferior compared to ours in pose error and expression error, revealing the importance of the reasonable space choice. Also, we can see that *$\mathcal{W}$ space* performs best in PSNR and SSIM, that is because face swapping in $\mathcal{W}$ space tends

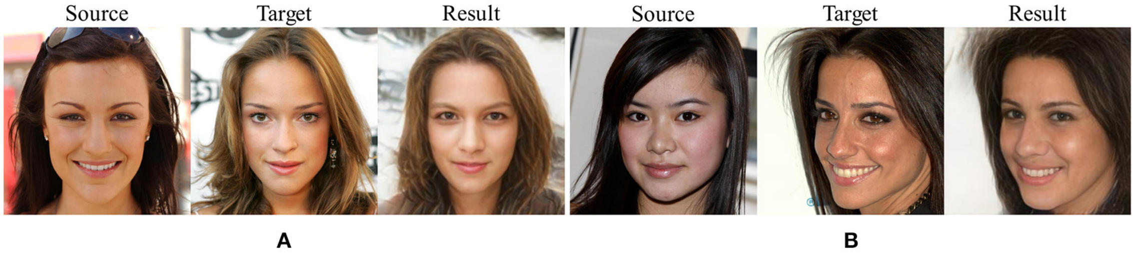


FIGURE 5 | Shortcomings of the proposed model. The problem in panel **(A)** is that the background of the conversion result is blurred. The problem in panel **(B)** is that the swapped face lacks Asian characteristics.

to map a wild face to a most similar face in the StyleGAN face domain, which is a more natural result with better image quality. Thanks to the help of StyleGAN, every model in **Table 3** surpasses the existing face-swapping methods in PSNR and SSIM.

4.4. Discussion

The core of the proposed model is to use StyleGAN as the face decoder, which reduces the burden of face spatial feature learning and dramatically reduces the possibility of artifacts in the conversion results. However, the proposed method also has some defects. As shown in **Figure 5A**, the letters in the background of the target image become blurred in the resulting image, which shows that the proposed model is not good at restoring the background. Although the pretrained model we use learns the potential features of face space, it does not learn well how to separate the head from the background. To deal with this problem, we will separate the head and background in the next step through image segmentation and then combine the background of the target image with the head of the resulting image. At the same time, **Figure 5B** shows that the resulting image lacks Asian characteristics similar to those in the source image, which reflects the problem of insufficient potential vectors in the StyleGAN face space and is caused by the relative lack of Asian faces in the training dataset. Therefore, adding more types of faces to the pretrained model and selecting a better-pretrained model should also be a focus in future work.

5. CONCLUSION

This article proposes a new face-swapping framework that includes ShapeEditor and a pretrained StyleGAN model. The pretrained model gives the proposed framework the potential to generate clear and realistic faces. The ShapeEditor encoder effectively extracts and integrates the attribute and identity information of the input images, then accurately maps them onto the $\mathcal{W}+$ space, thus controlling StyleGAN to output the appropriate results. Extensive experiments show that the proposed method performs better than existing frameworks in terms of clarity and authenticity, with sufficiently integrating identity and attribute.

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

SY is responsible for code writing and thesis writing. KQ, RQ, PX, and SS are responsible for the inspiration of ideas. NL, LW, JC, GH, and BY put forward their opinions on the revision of the paper. All authors contributed to the article and approved the submitted version.

REFERENCES

Abdal, R., Qin, Y., and Wonka, P. (2019). "Image2stylegan: how to embed images into the stylegan latent space?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 4432–4441.

Abirami, R. N., and Vincent, P. D. R. (2021). Identity preserving multi-pose facial expression recognition using fine tuned vgg on the latent space vector of generative adversarial network. *Math. Biosci. Eng.* 18, 3699–3717. doi: 10.3934/mbe.2021186

Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. (2018). "Towards open-set identity preserving face synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT) 6713–6722.

Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., and Nayar, S. K. (2008). Face swapping: automatically replacing faces in photographs. *ACM Trans. Graph.* 27, 39. doi: 10.1145/1360612.1360638

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). "Arcface: additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 4690–4699.

Gu, J., Shen, Y., and Zhou, B. (2020). "Image processing using multi-code gan prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 3012–3021.

Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., and Li, S. Z. (2020). "Towards fast, accurate and stable 3d dense face aliganment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16* (Glasgow: Springer), 152–168.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. (2020). Ganspace: aiscovering interpretable gan controls. *arXiv preprint* arXiv:2004.02546.

Huang, X., and Belongie, S. (2017). "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 1501–1510.

Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., and Huang, F. (2020). Curricularface: adaptive curriculum learning loss for deep face recognition. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 5901–5910.

Huynh-Thu, Q., and Ghanbari, M. (2008). Scope of validity of psnr in image/video quality assessment. *Electron. Lett.* 44, 800–801. doi: 10.1049/el:20080522

Karras, T., Laine, S., and Aila, T. (2019). "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 4401–4410.

Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 3677–3685.

Lee, C.-H., Liu, Z., Wu, L., and Luo, P. (2020). "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 5549–5558.

Li, L., Bao, J., Yang, H., Chen, D., and Wen, F. (2019). Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint* arXiv:1912.13457.

Li, Y., and Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint* arXiv:1811.00656.

Nandhini Abirami, R., Durai Raj Vincent, P., Srinivasan, K., Tariq, U., and Chang, C.-Y. (2021). Deep cnn and deep gan in computational visual perception-driven image analysis. *Complexity* 2021:5541134. doi: 10.1155/2021/5541134

Natsume, R., Yatagawa, T., and Morishima, S. (2018a). "Fsnet: an identity-aware generative model for image-based face swapping," in *Asian Conference on Computer Vision* (Perth: Springer), 117–132.

Natsume, R., Yatagawa, T., and Morishima, S. (2018b). Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint* arXiv:1804.03447.

Nirkin, Y., Keller, Y., and Hassner, T. (2019). "Fsgan: subject agnostic face swapping and reenactment," In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 7184–7193.

Nirkin, Y., Masi, I., Tuan, A. T., Hassner, T., and Medioni, G. (2018). "On face segmentation, face swapping, and face perception," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an: IEEE), 98–105.

Nitzan, Y., Bermano, A., Li, Y., and Cohen-Or, D. (2020). Face identity disentanglement via latent space mapping. *ACM Trans. Graph.* 39, 1–14. doi: 10.1145/3414685.3417826

Olszewski, K., Li, Z., Yang, C., Zhou, Y., Yu, R., Huang, Z., et al. (2017). "Realistic dynamic facial textures from a single image using gans," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 5429–5438.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 2337–2346.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., and Cohen-Or, D. (2020). Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint* arXiv:2008.00951.

Ross, A., and Othman, A. (2010). Visual cryptography for biometric privacy. *IEEE Trans. Inform. Forensics Secur.* 6, 70–81. doi: 10.1109/TIFS.2010.2097252

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). "FaceForensics++: learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)* (Seoul).

Shen, Y., Yang, C., Tang, X., and Zhou, B. (2020). Interfacegan: interpreting the disentangled face representation learned by gans. *IEEE Trans. Pattern Anal. Mach. Intell.* doi: 10.1109/TPAMI.2020.3034267

Shen, Y., and Zhou, B. (2020). Closed-form factorization of latent semantics in gans. *arXiv preprint* arXiv:2007.06600.

Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., and Schiele, B. (2018). "A hybrid model for identity obfuscation by face replacement," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 553–569.

Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.-P., Pérez, P., et al. (2020). "Stylerig: Rigging stylegan for 3d control over portrait images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 6142–6151.

Wang, X., Li, Y., Zhang, H., and Shan, Y. (2021). Towards real-world blind face restoration with generative facial prior. *arXiv preprint* arXiv:2101.04061.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Yao, G., Yuan, Y., Shao, T., and Zhou, K. (2020). "Mesh guided one-shot face reenactment using graph convolutional networks," in *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA), 1773–1781.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 586–595.

Zhao, Z., Liu, Q., and Zhou, F. (2021). "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 3510–3519.

Zhu, J., Shen, Y., Zhao, D., and Zhou, B. (2020). "In-domain gan inversion for real image editing," in *European Conference on Computer Vision* (Glasgow: Springer), 592–608.

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.

ORIGINAL RESEARCH
published: 28 February 2022
doi: 10.3389/fnbot.2022.827913

Adaptive Fusion Based Method for Imbalanced Data Classification

Zefeng Liang[1†], *Huan Wang*[2*†], *Kaixiang Yang*[3*] *and Yifan Shi*[4]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, [2] Guangdong Institute of Scientific and Technical Information, Guangzhou, China, [3] State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China, [4] College of Engineering, Huaqiao University, Quanzhou, China

The imbalance problem is widespread in real-world applications. When training a classifier on the imbalance datasets, the classifier is hard to learn an appropriate decision boundary, which causes unsatisfying classification performance. To deal with the imbalance problem, various ensemble algorithms are proposed. However, conventional ensemble algorithms do not consider exploring an effective feature space to further improve the performance. In addition, they treat the base classifiers equally and ignore the different contributions of each base classifier to the ensemble result. In order to address these problems, we propose a novel ensemble algorithm that combines effective data transformation and an adaptive weighted voting scheme. First, we utilize modified metric learning to obtain an effective feature space based on imbalanced data. Next, the base classifiers are assigned different weights adaptively. The experiments on multiple imbalanced datasets, including images and biomedical datasets verify the superiority of our proposed ensemble algorithm.

Keywords: imbalance learning, metric learning, information fusion, classification, ensemble learning

OPEN ACCESS

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Liang Liao,
Zhongyuan University of Technology, China
Peng Cao,
Northeastern University, China
Goksu Tuysuzoglu,
Dokuz Eylül University, Turkey

****Correspondence:***
Huan Wang
wanghuan6@email.szu.edu.cn
Kaixiang Yang
yangkaixiang@zju.edu.cn

†These authors have contributed equally to this work

Received: *02 December 2021*
Accepted: *25 January 2022*
Published: *28 February 2022*

Citation:
Liang Z, Wang H, Yang K and Shi Y (2022) Adaptive Fusion Based Method for Imbalanced Data Classification. Front. Neurorobot. 16:827913. doi: 10.3389/fnbot.2022.827913

1. INTRODUCTION

Many applications face imbalance problems (Farrand et al., 2020; Khushi et al., 2021; Zhang et al., 2021). The imbalance problem is caused by the difference in the number of samples in each class. When the classifiers are trained on imbalanced datasets, the classifiers tend to favor the majority class and predict more samples to be the majority class. Therefore, the minority class samples can not be correctly classified, which is called the imbalance problem. The imbalance problem is widespread in the applications, so more and more researchers focus on dealing with the imbalance problem.

To solve the imbalance problem, researchers have proposed various methods from different perspectives. Cost-sensitive method (Elkan, 2001) is a typical one. The cost-sensitive method assigns different classification losses to each class. The minority class has a higher classification loss than the majority class, such that the classifiers pay more attention to the minority class and get a correct result. Resampling is another typical method. Resampling methods remove or synthesize samples from the original data to balance the number of samples in each class, including undersampling, oversampling, and hybrid sampling. Undersampling (He and Garcia, 2009) method removes the majority class samples by some informed rules. Undersampling can produce a more clear decision boundary while the information of the excluded samples is lost.

On the other hand, the oversampling method proposes to generate the synthesis of minority class samples until the data is balanced. The synthesis samples may lie in the overlapped area and make the distribution worsen. To overcome the disadvantages, the hybrid sampling is investigated.

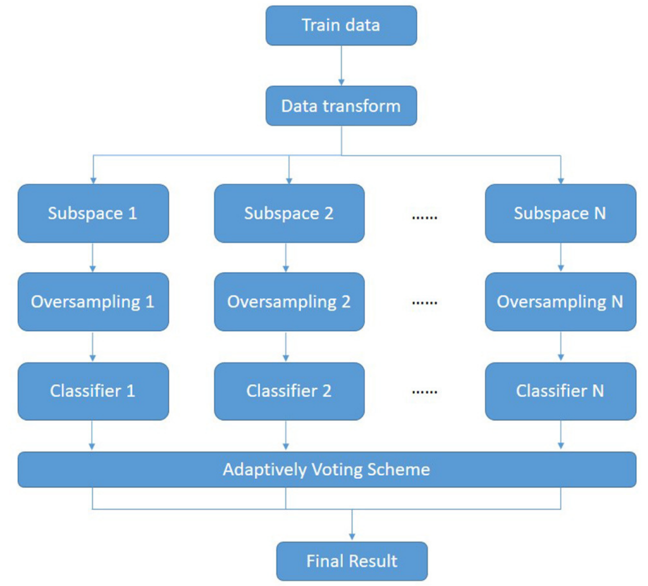


FIGURE 1 | The overall framework of ensemble algorithm.

As a two-stage strategy, hybrid sampling method combines undersampling and oversampling (I., 1976; Han et al., 2005). Other researchers combine the clustering method with oversampling (Barua et al., 2011).

Ensemble learning is widely used in solving the imbalance problem. Ensemble learning trains different classifiers and gets the result by integrated voting, which contains boosting and bagging (Ho, 1998; Skurichina and Duin, 2002; Chawla et al., 2003; Wang and Yao, 2009; Chen et al., 2010). Ensemble learning plays an essential role in the imbalance classification tasks (Bi and Zhang, 2018).

Nevertheless, the traditional imbalance algorithms have the following problems. Most algorithms do not consider mapping the imbalance data to another feature space for better classification performance. In addition, the importance of base classifiers is different, so it is inappropriate to treat the voting weight of base classifiers equally.

Metric learning is a hot topic in machine learning, which has been utilized in practical applications (Cao et al., 2019; Bai et al., 2021). Metric learning learns a feature space that is more effective than the original space. Euclidean distance is a common measure. However, Euclidean distance can not reflect the relationship correctly in the overlapped area. Consequently, some researchers capture the distance between samples by finding a transformation that can increase the distance between dissimilar samples and reduce the distance of similar samples (Köstinger et al., 2012). When training on the imbalance datasets, metric learning also suffers from imbalance problems (Gautheron et al., 2019). It needs to be modified before training on the imbalance datasets.

In this article, we propose an ensemble learning framework that combines metric learning and resampling. The metric learning is employed by building a feature space from the imbalanced dataset. The classifier is trained on the balanced datasets after oversampling on the feature space to reduce the impact of imbalance. Finally, the classifier is integrated by adaptive weighted voting.

The contributions of the article are as follows:

1) An imbalanced version of the large margin nearest neighbor (LMNN) algorithm is proposed to alleviate the influence of imbalanced data distribution and learn a robust feature space.

2) An GA-based weighting scheme is designed to adaptively optimize the importance of different classifiers.

3) Extensive experiments are conducted on various imbalanced datasets to verify the effectiveness of the proposed approach.

The main framework is as follows: Section 2 introduces the related work about resampling, ensemble learning, and metric learning; Section 3 discusses the proposed ensemble framework in detail; Section 4 shows the experiments about our proposed methods and discusses the result of the experiments. Section 5 draws the conclusion and future study.

2. RELATED WORK

The resampling method contains undersampling and oversampling. To get a balanced dataset, undersampling methods remove the majority of samples randomly or by informed rules. Tomek link (Batista et al., 2004) removes samples that are of a different class from the neighbor. Undersampling can reduce the imbalance problem, while it may suffer from information loss. When the number of samples in each class is quite different, most of the majority of class samples are removed. Hence, the information of the majority of class is lost severely. On the other hand, oversampling proposes to generate synthesis minority class samples to balance the dataset. SMOTE (Chawla et al., 2002) propose to generate samples by interpolating between a given sample and its neighbors. The synthesis sample x_i^* is generated as follows:

$$x_i^* = x_i + (x_n - x_i) * r \tag{1}$$

In which x_n is the neighbor of sample x_i and r is a random value between $[0, 1]$. Adaptive Synthetic sampling approach (ADASYN) (He et al., 2008) makes different samples generate different numbers of synthetic samples. Some methods combine oversampling with clustering to overcome the problem that synthesized samples are located in the overlapped area. Majority Weighted Minority Oversampling Technique (MWMOTE) (Barua et al., 2014) generates minority samples within the cluster. Additionally, geometric-SMOTE (Douzas and Bacao, 2019) proposes a universal method that can be used in most oversampling methods. Mahalanobis Distance-based Oversampling technique (MDO) (Abdi and Hashemi, 2016) and its variant (Yang et al., 2018) propose to generate samples in the principal component space.

Euclidean distance is a traditional measure to reflect the similarity between samples. However, dissimilar samples may be closer to the similar samples in the overlapped area, which is inefficient to apply Euclidean distance. Metric learning learns a feature space that can reflect the relationship between samples more correctly. In the feature space, similar samples are closer while dissimilar samples are separated apart. To achieve this

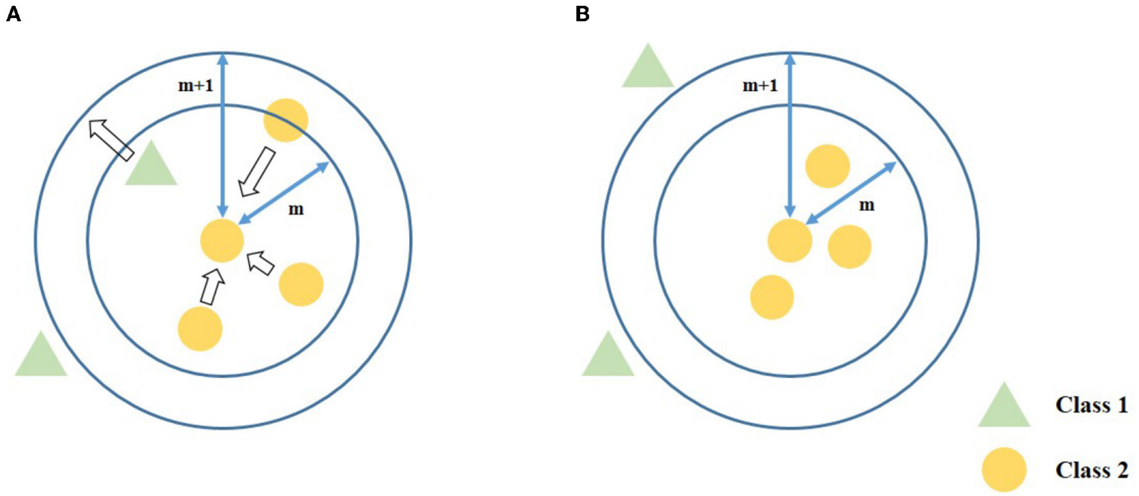


FIGURE 2 | The diagram of proposed data transformation. After the data transformation, the similar neighbors are closer, while the dissimilar neighbor samples are pushed and hold a certain distance to the anchor sample. **(A)** Before data transformation. **(B)** After data transformation.

goal, many metric learning algorithms have been proposed. LMNN (Weinberger and Saul, 2009) minimizes the distance between the anchor sample and its neighbors of the same class. At the same time, the anchor sample maintains a margin with neighbors of a different class. Information-theoretic metric learning (ITML) (Davis et al., 2007) makes distribution on the feature space similar to the Gaussian distribution. Some methods utilize metric learning on imbalanced datasets. Imbalance metric learning (IML) (Gautheron et al., 2019) modified LMNN by assigning different weights to sample pairs. Distance Metric by Balancing KL-divergence (DMBK) (Feng et al., 2019) balances the divergence of each class on the feature space. Iterative metric learning (Wang et al., 2018) learns a feature space for the area near each testing data.

Ensemble learning integrates classifiers to improve the robustness and performance of classification results. EasyEnsemble (Liu et al., 2009) trains several classifiers on the subset, which contains part of majority class samples and whole minority class samples. BalanceCascade (Liu et al., 2009) splits the majority class samples as several subsets and trains AdaBoost classifiers based on the subsets. Yang et al. (2021) proposes an ensemble framework based on subspace feature space ensemble and metric learning.

3. PROPOSED METHODOLOGY

In this section, we propose an ensemble framework combining metric learning with oversampling. **Figure 1** shows the overall framework of our proposed algorithm. First, the metric learning methods based on the imbalance problem(denoted as ImLMNN) are applied for getting a better feature space L. The data X is transformed by mapping matrix L and gets the mapped data X^*. Then, the feature space S_i is constructed. Next, the oversampling method is employed for getting a balance training dataset S_i^*. Finally, different classifiers are applied in balance datasets and voting for the result. The pseudo-code is shown in the **Algorithm 1**.

TABLE 1 | The attributes of datasets.

	IR	Samples	Features
climate	10.74	540	18
libras_move	14.00	360	90
ecoli2	5.46	90	7
glass_0_1_2_3_vs_4_5_6	3.20	214	9
yeast3	8.10	1484	8
cleveland_0_vs_4	12.31	173	13
winequality_red_4	29.17	1599	11
ecoli1	3.36	336	7

We aim at finding a feature space that can better describe the sample relationship to improve the performance of classifiers. LMNN transforms data to a latent feature space, in which similar samples are closer while dissimilar samples are separated apart. The loss function of LMNN is as follows:

$$f(\mathbf{L}) = f_{push}(\mathbf{L}) + f_{pull}(\mathbf{L}) \tag{2}$$

where

$$f_{pull}(\mathbf{L}) = \sum_{i,j} \left\| \mathbf{L}\left(x_i - x_j\right) \right\|^2 \tag{3}$$

$$f_{push}(\mathbf{L}) = \sum_{i,j} \sum_{l} \left(1 - y_{il}\right) \left[1 + \left\| \mathbf{L}\left(x_i - x_j\right) \right\|^2 - \left\| \mathbf{L}\left(x_i - x_l\right) \right\|^2 \right]_+ \tag{4}$$

The loss function of LMNN contains $f_{push}(\mathbf{L})$ and $f_{pull}(\mathbf{L})$. $f_{pull}(\mathbf{L})$ reduces the distance between the anchor and its similar neighbors, while $f_{push}(\mathbf{L})$ penalizes the distance between

TABLE 2 | Comparisons between imbalance learning algorithm and our proposed method in terms of AUC.

	SMOTE	RandomForest	RUSboost	Balance bagging	LMNN ensemble	ImLMNN ensemble
climate	0.8785 ± 0.0092	0.5353 ± 0.0188	0.7462 ± 0.0245	0.8496 ± 0.0206	0.8787 ± 0.0162	**0.8796 ± 0.021**
libras_move	0.8741 ± 0.0227	0.8133 ± 0.025	0.8015 ± 0.06	0.8544 ± 0.0153	0.8146 ± 0.0286	**0.8975 ± 0.0144**
ecoli2	0.8671 ± 0.0121	0.813 ± 0.0113	0.8252 ± 0.0504	0.8678 ± 0.0138	0.8537 ± 0.0102	**0.873 ± 0.0086**
glass_0_1_2_3_vs_4_5_6	0.8945 ± 0.0172	0.876 ± 0.0129	0.8441 ± 0.039	0.8804 ± 0.0251	0.827 ± 0.0295	**0.9064 ± 0.0173**
yeast3	0.8278 ± 0.8278	0.7348 ± 0.0219	0.8163 ± 0.0258	0.8256 ± 0.0273	0.8381 ± 0.0163	**0.8395 ± 0.0125**
cleveland_0_vs_4	0.8919 ± 0.0103	0.6367 ± 0.088	0.7547 ± 0.05	0.848 ± 0.0405	0.8871 ± 0.0236	**0.8955 ± 0.0419**
winequality_red_4	0.6667 ± 0.0175	0.5291 ± 0.0093	0.5919 ± 0.0323	0.6515 ± 0.6515	0.6615 ± 0.0111	**0.6776 ± 0.0116**
ecoli1	0.8715 ± 0.0208	0.8461 ± 0.0297	0.7868 ± 0.0464	0.8756 ± 0.0228	0.8786 ± 0.0062	**0.8804 ± 0.0117**
AVERAGE_AUC	0.8465	0.723	0.7708	0.8316	0.8299	**0.8562**

The bold value means the best result among the compared algorithms.

Algorithm 1 : Imbalance Ensemble Framework.

```
Input: Training set $\mathbf{X} = (x_1, x_2, \ldots, x_n)$
Parameter: Number of subspace $N$
Procedure:
Obtain the feature space $\mathbf{L}$ by ImLMNN;
Map the data by learned feature space $\mathbf{L}$ and get the mapped dataset $\mathbf{X}^*$;
for $i$ in 1,..., $N$ do
    Extract part of features by a threshold $M$ and get the subspace $\mathbf{S}_i$;
    Apply SMOTE oversampling method on subspace $\mathbf{S}_i$ and get the balanced subset $\mathbf{S}_i^*$;
    Classify on subset $\mathbf{S}_i^*$ and get predict result $y_i^*$;
end for
Vote for result by adaptive weight $\mathbf{W} = [w_1, w_2, ..., w_N]$;
Output: The final result $\mathbf{Y} = \left(y^1_{final}, y^2_{final}, \ldots, y^n_{final}\right)$.
```

dissimilar samples. $[z]_+ = \max(z, 0)$ is the hinge loss. $y_{il} = 1$ when x_i and x_l belong to the same class, otherwise $y_{il} = 0$.

However, the LMNN algorithm is inappropriate directly to apply in imbalanced datasets. To solve this problem, we assign different weights to samples. The weight w_i of sample x_i is as follows:

$$w_i = \frac{\delta^i}{|N_c| * d\left(x_i, \overline{X_c}\right)} \tag{5}$$

The loss of samples is divided by the number of samples N_c in the corresponding class, such that the impact caused by the imbalance problem is alleviated. To emphasize the samples near decision boundaries, we compute the sum of the density of majority class δ_n^i and minority class δ_p^i as density δ^i. The density δ^i is defined as follows:

$$\delta^i = \delta_n^i + \delta_p^i \tag{6}$$

$$\delta_n^i = \frac{1}{\frac{1}{k}\sum\limits_{j=1}^{k} d_{ij}}, \quad \delta_p^i = \frac{1}{\frac{1}{h}\sum\limits_{j=1}^{h} d_{ij}} \tag{7}$$

δ_n^i and δ_p^i describe the aggregation of samples in neighboring areas about majority class and minority class. k and h are the number of neighbor samples in calculating δ_n^i and δ_p^i, respectively. When the density δ_c^i is large, the samples in class c are close to x_i. Therefore, a large sum of density δ^i reflects that sample x_i is close to samples of both majority and minority classes or in the inner of class with high density.

Outliers and noises are also in the border area. To alleviate the influence of outliers and noises, we divide sample weight by $d(x_i, \bar{X}_c)$ which is the distance between sample x_i and the center of class c. The center of class c is defined as:

$$\overline{X_c} = \sum_{\{i|y_i=c\}} x_i. \tag{8}$$

Therefore, the overall objective function of the data transformation algorithm is:

$$f(\mathbf{L}) = \sum_{i,j} w_i \left\|\mathbf{L}\left(x_i - x_j\right)\right\|^2 + \sum_{i,j}\sum_{i} w_i\left(1 - y_{il}\right)\left[1 + \left\|\mathbf{L}\left(x_i - x_j\right)\right\|^2 - \left\|\mathbf{L}\left(x_i - x_l\right)\right\|^2\right]_+ \tag{9}$$

The diagram of data transformation is shown in **Figure 2**. Similar to LMNN, Equation (9) contains f_{pull} which pulls similar samples closer and f_{push} which push dissimilar samples separate apart. In addition, each sample has a different weight to deal with the imbalance problem.

After the data transformation, we extract M features to build the subspace $\mathbf{S}_i$:

$$\mathbf{S}_i = [f_i^1, f_i^2, ..., f_i^M] \tag{10}$$

The feature is extracted N times to generate N subspace. In the subspace, the dataset is still imbalanced, which affects the classifier's performance. To solve this problem, oversampling is utilized in each feature subspace. Specifically, the subset is $\mathbf{S}_i^* = \mathbf{S}_i \cup \mathbf{Syn}_i$. The $\mathbf{Syn}_i$ is the synthesis minority data that is generated by SMOTE on feature subspace $\mathbf{S}_i$ and helps to form a new balanced subset $\mathbf{S}_i^*$ with the original subspace.

The classifier c_i is trained on the balance subset $\mathbf{S}_i^*$ and votes for obtaining the result. However, the performance of each

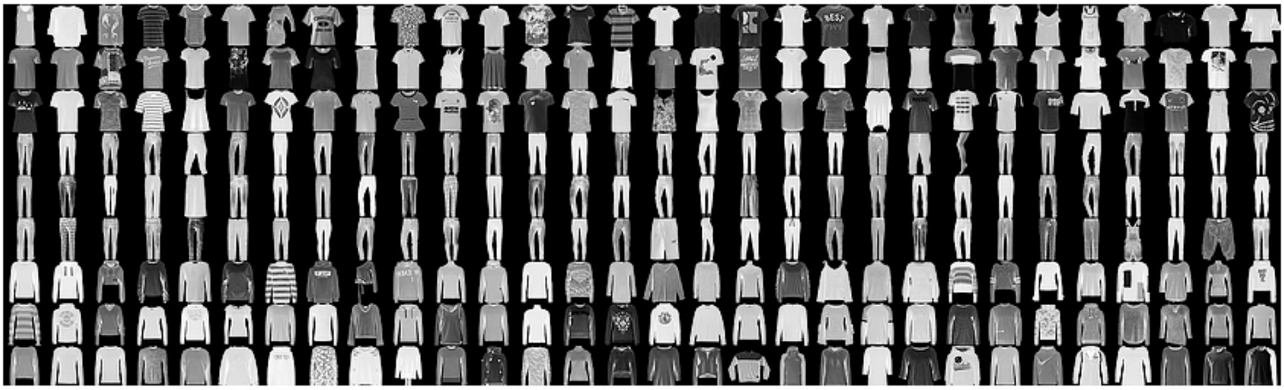


FIGURE 3 | The Fashion-mnist dataset.

TABLE 3 | Comparisons between imbalance learning algorithm and our proposed method in terms of AUC.

	SMOTE	RandomForest	RUSboost	Balance bagging	LMNN ensemble	ImLMNN ensemble
Fashion-mnist	0.9384 ± 0.0018	0.9523 ± 0.0025	0.9414 ± 0.006	0.9561 ± 0.0043	0.9543 ± 0.0044	**0.9610 ± 0.0012**

The bold value means the best result among the compared algorithms.

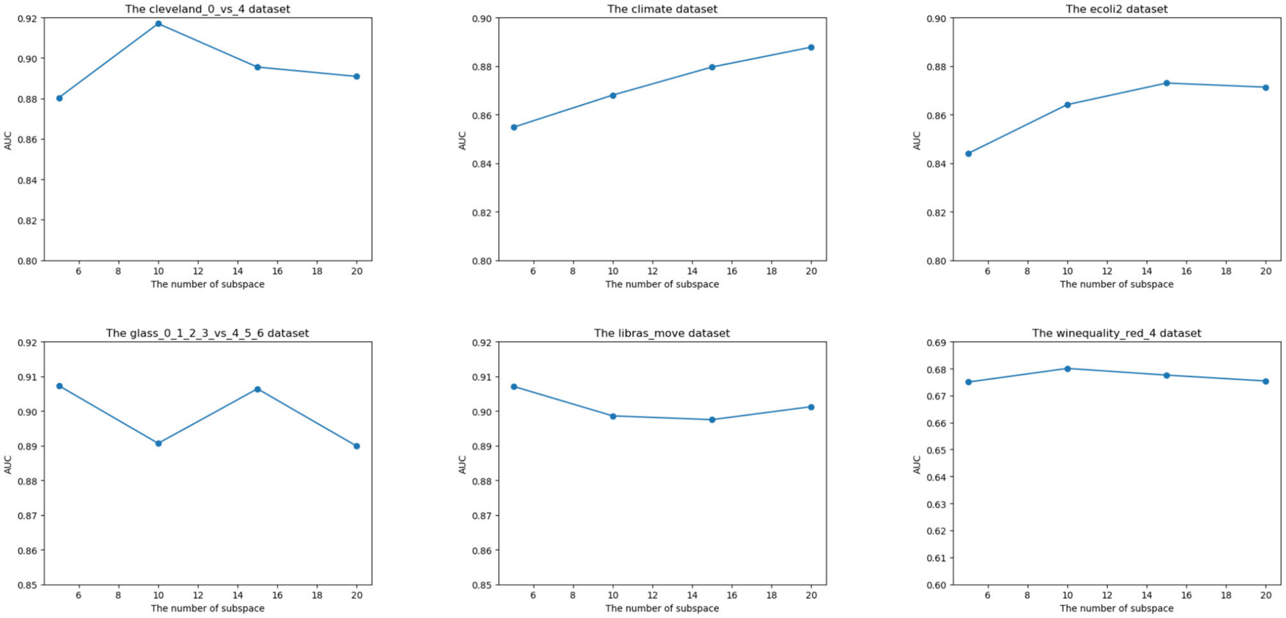


FIGURE 4 | Effect of the number of subspaces on the performance of ensemble on six datasets.

classifier is different. The contribution of each classifier in voting should be determined by the classification result, rather than being treated equally. Therefore, we utilize the weight assign process to progressively vote. In detail, the GA algorithm is applied to obtain the weight of each classifier adaptively. The detailed description is shown as **Algorithm 2**.

First, the initial genes $\mathbf{G} = (g_1, g_2, ...g_n)$ is generated as the weight of the classifier, in which n is the number of subspace and g_i is the weight of classifier i. Next, the GA algorithm finds the n individual and does crossover and mutation. Given two parent genes $g_i = [p_i^1, p_i^2, ..., p_i^S]$ and $g_j = [p_j^1, p_j^2, ..., p_j^S]$ with length S, the crossover method exchanges part of features in genes. Suppose the exchange occurs at position α ($\alpha \in [1, S]$), then the genes after the exchange are:

$$\begin{aligned} g_i^* &= [p_i^1, p_i^2, ..., p_j^\alpha, ..., p_i^S] \\ g_j^* &= [p_j^1, p_j^2, ..., p_i^\alpha, ..., p_j^S] \end{aligned} \tag{11}$$

Algorithm 2 : Adaptive weight Procedure.

```
Input: Classifier set $\mathbf{C} = (c_1, c_2, \ldots, c_n)$
Parameter: Number of genes $n$, Population size of genes $n_p$, Max iteration $N$
Initialize: genes $\mathbf{G} = \left(g_1, g_2, \ldots, g_n\right)$
Procedure:
while not converge do
    Select genes as parent randomly;
    Do crossover and mutation on parent genes to generate $n_{child}$ child genes by Eq. (11) and Eq. (12);
    for $i$ in 1,...,$n$ do
        Calculates the fitness of gene $Fparent_i$;
    end for
    for $j$ in 1,...,$n_{child}$ do
        Calculates the fitness of gene $Fchild_j$;
    end for
    if $Fchild_h > Fparent_l$ then
        replace parent gene $g_l$ by child gene $g_h$;
    end if
    Until $N$ Converge
end while
Output: The optimal genes $\mathbf{G}^* = \left(g_1^*, g_2^*, \ldots, g_n^*\right)$.
```

The mutation may occur in each position of genes. Suppose the mutation happens in position γ ($\gamma \in [1, S]$) of gene g_k, then we have:

$$g_k^* = [p_k^1, p_k^2, ..., p_k^{\gamma}, ..., p_k^S] \tag{12}$$

in which p_k^{γ} is a random value. After crossover and mutation, the child's genes are generated. We calculate the fitness of parent genes $Fparent_l$ ($l \in [1, n]$) and child genes $Fchild_h$ ($h \in [1, n_{child}]$). The fitness is set as AUC value. Finally, the parent genes are replaced by child genes with higher fitness values. When the iteration is over, the optimal classifier weight $\mathbf{G}^* = (g_1^*, g_2^*, ..., g_n^*)$ is obtained.

We can get the final result by weighted voting integration. The result of classifier c_i is denoted as y_i. Then, the final result is

$$y_{final} = \sum_N g_i^* y_i \tag{13}$$

4. EXPERIMENT

In this section, we show the experiments about the proposed ensemble framework and compare the algorithm on various datasets from UCI (Dua and Graff, 2017) and KEEL (Alcala-Fdez et al., 2010). Our algorithm is also applied in the Fashion-mnist image dataset. Finally, we analyze the effect of parameters on our proposed algorithm.

4.1. Datasets

To evaluate the performance of our algorithm, we choose eight datasets from UCI and KEEL with different attributes, such as imbalance ratio (IR), number of samples, and features. The attributes are shown in **Table 1** in detail.

4.2. Evaluation Criteria

Accuracy is the typical criterion to evaluate the performance of the algorithm. However, due to the imbalance problem, accuracy is inappropriate for imbalance learning. AUC (Fawcett, 2004) is the area under the receiver operating characteristic curve, which is not sensitive to the imbalance data, and it is widely used in imbalance learning. In the experiments, we use AUC as evaluation criteria.

4.3. Comparison With Other Algorithms

To show the superiority, several algorithms and imbalance ensemble frameworks are compared with our algorithm. Specifically, we choose RandomForest, RUSboost, and BalanceBagging as the baseline. In addition, SMOTE algorithm is also chosen. The baseline algorithms are shown as follows:

1. SMOTE: A typical oversampling method. It generates samples by interpolating between samples and their neighbors.
2. RandomForest: An ensemble framework that uses bagging to build subsets for tree classifiers. The number of trees we set is 15.
3. RUSboost: A hybrid method that combines sampling with boosting. The number of iterations is 15.
4. BalanceBagging: A variant of Bagging that is applied sampling in each bootstrap. The number of subspaces we set is 15.

For our proposed algorithm, the number of subspaces is 15, and the ratio of extracted features is 0.7. To show the ablation experiment, we compare the LMNN ensemble, which replaces our proposed data transformation algorithm ImLMNN with the original LMNN algorithm. We choose linear SVM as the base classifier. The algorithms run five times and calculate average AUC as evaluation criteria. The 5-fold cross-validation is also applied. The result of the experiment is shown in **Table 2**.

From **Table 2**, we can see that our algorithm has the highest average AUC on given datasets, which is superior to other compared algorithms. Compared with other algorithms, the proposed algorithm has at least a 1% improvement in average AUC. Also, compared with the ensemble algorithm that applied the original LMNN algorithm as data transformation, our proposed method has a near 3% improvement in average AUC. Our method takes data transformation in the imbalanced datasets into account, which is superior to other compared algorithms.

4.4. Comparison With Different Algorithms on Image Dataset

Our algorithm is applied in the image dataset and compared with other algorithms. **Figure 3** shows part of samples in the Fashion-mnist dataset. Fashion-mnist is a famous image dataset that has 784 features and 60,000 samples. The dataset has 10 classes. To build the imbalanced dataset, we choose the T-shirt class as the majority class and the pullover class as the minority class. The majority and minority class samples are 3,000 and 600, respectively, which means the imbalance ratio is 5:1. Considering that the feature size is similar to the number of samples, we set the feature extraction ratio to 0.1. **Table 3** shows the result of the experiment.

We can see that our method also has the best performance in the Fashion-mnist dataset. The reason is that our proposed method can deal with the imbalance problem in the image dataset.

4.5. The Effect of Parameter

In this section, we show the impact of the parameter in our algorithm. The number of subspaces influences the performance of our ensemble framework. We set the number of subspace to [5,10,15,20]. The parameter experiment result is shown in **Figure 4**.

For Cleveland_0_vs_4, ecoli2, and winequality_red_4 datasets, the AUC is improved when the value of subspace is increased. However, when the number of subspaces exceeds a specific number, the AUC decreases as the subspace increases. The reason is that the increasing value of subspaces improves the diversity of subspace, while the excessive subspaces introduce redundant information and are harmful to the algorithm's performance. For other datasets, the trend of AUC is diverse due to the uncertainty of the GA algorithm. Considering the algorithm result on the overall dataset, the proposed number of the subspace is 15.

5. CONCLUSION AND FUTURE WORK

In this article, we propose an ensemble framework to deal with imbalanced datasets. We explore an effective feature space to improve the performance of the subsequent procedure. In addition, we propose an adaptive integrated voting process to assign weights for classifiers. The experiments on various real-world imbalanced datasets, including the imbalanced image dataset, show the superiority of the proposed ensemble framework. Finally, we show the experiment to explore the effect of the parameter.

Future study contains several points: (1) Various methods can transform data into other feature spaces, so choosing appropriate methods should be considered. (2) A more effective adaptive weight process should be explored to assign weight based on the performance of the base classifiers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://archive.ics.uci.edu/ml/index.php.

AUTHOR CONTRIBUTIONS

ZL, HW, and KY contributed to conception and design of the study. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The work described in this study was partially funded by the National Key Research and Development Program of China under Grant 2019YFB1703600, in part by the National Natural Science Foundation of China No. 62106224 and Key-Area Research and Development Program of Guangdong Province (No. 2018B010107002).

REFERENCES

Abdi, L., and Hashemi, S. (2016). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Trans. Knowl. Data Eng.* 28, 238–251. doi: 10.1109/TKDE.2015.2458858

Alcala-Fdez, J., Fernández, A., Luengo, J., Derrac, J., Garc'ia, S., Sanchez, L., et al. (2010). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J. Multiple Valued Logic Soft Comput.* 17, 255–287.

Bai, S., Luo, Y., Yan, M., and Wan, Q. (2021). Distance metric learning for radio fingerprinting localization. *Exp. Syst. Appl.* 163:113747. doi: 10.3390/s21134605

Barua, S., Islam, M. M., and Murase, K. (2011). *A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning* (Shanghai: Springer), 735–744.

Barua, S., Islam, M. M., Yao, X., and Murase, K. (2014). MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng.* 26, 405–425. doi: 10.1109/TKDE.2012.232

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* 6, 20–29. doi: 10.1145/1007730.1007735

Bi, J., and Zhang, C. (2018). An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *Knowl. Based Syst.* 158, 81–93. doi: 10.1016/j.knosys.2018.05.037

Cao, X., Ge, Y., Li, R., Zhao, J., and Jiao, L. (2019). Hyperspectral imagery classification with deep metric learning. *Neurocomputing* 356, 217–227. doi: 10.1016/j.neucom.2019.05.019

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1007/978-3-540-39804-2_12

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003). "Smoteboost: improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases: PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases* (Cavtat), 107–119.

Chen, S., He, H., and Garcia, E. A. (2010). Ramoboost: ranked minority oversampling in boosting. *IEEE Trans. Neural Netw.* 21, 1624–1642. doi: 10.1109/TNN.2010.2066988

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007). *Information-Theoretic Metric Learning* (Corvallis, OR: Association for Computing Machinery), 209–216.

Douzas, G., and Bacao, F. (2019). Geometric smote a geometrically enhanced drop-in replacement for smote. *Inf. Sci.* 501, 118–135. doi: 10.1016/j.ins.2019.06.007

Dua, D., and Graff, C. (2017). UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. Available online at: http://archive.ics.uci.edu/ml

Elkan, C. (2001). "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (San Francisco, CA) 973–978.

Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. (2020). "Neither private nor fair: impact of data imbalance on utility and fairness in differential privacy," in *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice PPMLP'20* (New York, NY: Association for Computing Machinery), 15–19.

Fawcett, T. (2004). Roc graphs: notes and practical considerations for data mining researchers. *ReCALL* 31, 1–38.

Feng, L., Wang, H., Jin, B., Li, H., Xue, M., and Wang, L. (2019). Learning a distance metric by balancing kl-divergence for imbalanced datasets. *IEEE Trans. Syst. Man Cybern. Syst.* 49, 2384–2395. doi: 10.1109/TSMC.2018.2790914

Gautheron, L., Habrard, A., Morvant, E., and Sebban, M. (2019). *Metric Learning From Imbalanced Data* (Portland, OR: Institute of Electrical and Electronics Engineers), 923–930.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). *Borderline-Smote: a New Over-Sampling Method in Imbalanced Data Sets Learning* (Hefei: Springer), 878–887.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong: IEEE), 1322–1328. doi: 10.1109/IJCNN.2008.4633969

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844.

I., T. (1976). Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.* SMC-6, 769–772.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., et al. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access* 9, 109960–109975. doi: 10.1109/ACCESS.2021.3102399

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI), 2288–2295.

Liu, X., Wu, J., and Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* 39, 539–550. doi: 10.1109/TSMCB.2008.2007853

Skurichina, M. and Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal. Appl.* 5, 121–135. doi: 10.1007/s100440200011

Wang, N., Zhao, X., Jiang, Y., Gao, Y., and BNRist, K. (2018). "Iterative metric learning for imbalance data classification," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (Stockholm), 2805–2811.

Wang, S. and Yao, X. (2009). "Diversity analysis on imbalanced data sets by using ensemble models," in *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009, part of the IEEE Symposium Serie on Computational Intelligence* (Nashville, TN), 324–331.

Weinberger, K. Q., and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10, 207–244.

Yang, K., Yu, Z., Chen, C. L. P., Cao, W., Wong, H.-S., You, J., and Han, G. (2021). Progressive hybrid classifier ensemble for imbalanced data. *IEEE Trans. Syst. Man Cybern. Syst.* 1–15. doi: 10.1109/TSMC.2021.3051138

Yang, X., Kuang, Q., Zhang, W., and Zhang, G. (2018). Amdo: An over-sampling technique for multi-class imbalanced problems. *IEEE Trans. Knowl. Data Eng.* 30, 1672–1685. doi: 10.1109/TKDE.2017.2761347

Zhang, J., Zhang, Q., He, X., Sun, G., and Zhou, D. (2021). Compound-fault diagnosis of rotating machinery: A fused imbalance learning method. *IEEE Trans. Control Syst. Technol.* 29, 1462–1474. doi: 10.1109/TCST.2020.3015514

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 08 March 2022
doi: 10.3389/fnbot.2022.846580

Multi-Exposure Image Fusion Algorithm Based on Improved Weight Function

Ke Xu, Qin Wang, Huangqing Xiao and Kelin Liu*

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

High-dynamic-range (HDR) image has a wide range of applications, but its access is limited. Multi-exposure image fusion techniques have been widely concerned because they can obtain images similar to HDR images. In order to solve the detail loss of multi-exposure image fusion (MEF) in image reconstruction process, exposure moderate evaluation and relative brightness are used as joint weight functions. On the basis of the existing Laplacian pyramid fusion algorithm, the improved weight function can capture the more accurate image details, thereby making the fused image more detailed. In 20 sets of multi-exposure image sequences, six multi-exposure image fusion methods are compared in both subjective and objective aspects. Both qualitative and quantitative performance analysis of experimental results confirm that the proposed multi-scale decomposition image fusion method can produce high-quality HDR images.

Keywords: high dynamic range image, multi-scale decomposition, multi-exposure images, image fusion, Laplacian pyramid (LP)

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Yu Liu,
Hefei University of Technology, China
Huafeng Li,
Kunming University of Science and Technology, China

Correspondence:
Ke Xu
xuke@cqupt.edu.cn

Received: *31 December 2021*
Accepted: *07 February 2022*
Published: *08 March 2022*

Citation:
Xu K, Wang Q, Xiao H and Liu K (2022) Multi-Exposure Image Fusion Algorithm Based on Improved Weight Function.
Front. Neurorobot. 16:846580.
doi: 10.3389/fnbot.2022.846580

1. INTRODUCTION

Due to the limited dynamic range of imaging equipment, it is impossible for existing imaging equipment to capture all the details in one scene with a single exposure. Therefore, underexposure or overexposure often occurs in daily shooting, which seriously affects the visualization of images and the display of key information. High-dynamic-range (HDR) imaging techniques overcome this limitation, but most of currently used standard monitors use low dynamic range (LDR) (Ma et al., 2015a). So, a tone mapping process is required to compress the dynamic range of HDR images for display after acquiring HDR images. Multi-exposure image fusion (MEF) methods use a cost-effective way to solve the dynamic range mismatch between HDR imaging and LDR display. Source image sequences with different exposure levels are taken as input and the brightness information in accordance with the human visual system (Ma et al., 2017) is fused with them to generate HDR images with rich information and sensitive perception.

In recent years, many MEF algorithms have been developed. Like multi-source image fusion (Jin et al., 2021a), MEF algorithms are usually divided into four categories (Liu et al., 2020): spatial domain methods, transform domain methods, the combination of spatial domain and transform domain methods and deep learning methods (Jin et al., 2021b). This article mainly studies the MEF method in spatial domain,these methods mainly focus on providing the weighted sum of the input exposures image to obtain the fused image. Different MEF methods use different techniques to obtain the suitable weight map. Li et al. (2013) obtained the corresponding base and detail layers by decomposing the source image in two scales, and then processed them separately to obtain the final fusion image. Liu and Wang (2015) applied dense scale invariant feature transform (SIFT) (Liu et al., 2016) to obtain both contrast and spatial consistency weights based on local gradient information. Mertens et al. (2010) applied multi-resolution exposure sequences to Laplacian pyramid-based image fusion. The weighted average value was first calculated from the weighted values determined by contrast, saturation and good exposure, and then applied to obtain the pyramid coefficients. Finally, image fusion was achieved by reconstructing the obtained pyramid coefficients. Shen et al. (2014) proposed an exposure fusion method based on hybrid exposure weights and an improved Laplacian pyramid. This method considers the gradient vectors between different exposure source images, and uses an improved Laplacian pyramid to decompose input signals into both base and detail layers. Shen et al. (2011) proposed a probability model of MEF. According to the two quality indicators of both local contrast and color consistency of source image sequences, the generalized random walk framework was first used to calculate the optimal probability set. Then, the obtained probability set was used as the corresponding weights to realize image Fusion. Fei et al. (2017) applied an image smoothing algorithm based on weighted least squares to MEF for achieving detail extraction of HDR scenes. The extracted detail information was used in the multi-scale exposure fusion algorithm to achieve image fusion. So, fused images with rich colors and detailed information can be obtained. Li and Kang (2012) proposed a fusion method based on weighted sum. Firstly, three image features composed of local contrast, brightness and color differences are measured to estimate the weight, and then the weight map is optimized by recursive filter. Zhang and Cham (2012) proposed a simple and effective method, which uses gradient information to complete multi exposure image synthesis in static and dynamic scenes. Given multiple images with different exposures, the proposed method can seamlessly synthesize them under the guidance of gradient based quality evaluation, so as to produce a pleasant tone mapped high dynamic range image. Ma et al. (2017) proposed a method based on image structure block decomposition, which represents the image block with average intensity, signal intensity and signal structure, and then uses the intensity and exposure factor of the image block for weighted fusion, which can be used for both static scene fusion and dynamic scene fusion. Moriyama et al. (2019) proposed to use the light conversion method of preserving hue and saturation to generate a new multi exposure image for fusion, realize brightness conversion based on local color correction, and obtain the fused image by weighted average (weight is calculated by saturation). Wang and Zhao (2020) proposed using the super-pixel segmentation method to divide the input image into non overlapping image blocks composed of pixels with similar visual attributes, decompose the image block into three independent components: signal intensity, image structure and intensity, and then fuse the three components respectively according to the characteristics of human visual system and the exposure level of the input image. Qi et al. (2020) used the exposure quality a priori to select the reference image, used the reference image to solve the ghosting problem in the dynamic scene in the structural consistency test, and then decomposed the image by using the guidance filter, and proposed a fusion method combining spatial domain scale decomposition, image block structure decomposition and moderate exposure evaluation. Li et al. (2020) proposed a multi exposure image fusion algorithm based on improved pyramid transform. The algorithm improves the local contrast information of the image by using the adaptive histogram equalization algorithm, and calculates the image fusion weight coefficient with good contrast information, image entropy and exposure. Hayat and Imran (2019) proposed a ghosting free multi exposure image fusion technology based on dense sift descriptor and guided filter. Ulucan et al. (2020) proposed a new, simple and effective still image exposure fusion method. This technique uses weight map extraction based on linear embedding and watershed masking. Xu et al. (2021). Proposed a new multi exposure image fusion method based on tensor product and tensor singular value decomposition. A new fusion strategy is designed by using tensor product and t-svd. The luminance and chrominance channels are fused respectively to maintain color consistency. Finally, the chrominance and luminance channels are fused to obtain the fused image.

Both multi-scale decomposition method and fusion strategy of multi-scale coefficients determine the performance of the image fusion framework based on multi-scale decomposition. Pyramid transformation is a commonly used multi-scale decomposition method. Due to different scales and resolutions, the corresponding decomposition layer has different image feature information. In addition, the weight function design of feature extraction plays a decisive role in the final fusion result. Therefore, this article, proposes a fast and effective image fusion method based on improved weight function. The fusion weight map is calculated through the evaluation of exposure moderation and relative brightness. Combined with pyramid multi-scale decomposition, images with different resolutions are fused to generate the required high dynamic range image.

The rest of this article is organized as follows. The second section describes the overall process of the fusion algorithm; The third section is a detailed explanation of the weight function; The fourth section describes the process of image Gaussian pyramid decomposition and Laplace pyramid decomposition; The fifth section is the experimental results and analysis; The sixth section is the summary of this article.

2. WORKFLOW OF IMAGE FUSION ALGORITHM

MEF aims to generate an image containing the best pixel information from a series of images with different exposure levels. The pixel-based MEF performs weighted image fusion as follows.

$$FI(x,y) = \sum_{n=1}^{N} W_n(x,y) I_n(x,y) \tag{1}$$

where FI represents the fusion image, (x, y) represents pixel coordinates, N represents the number of images, I_n represents the pixel intensity of the nth image, and W_n represents the

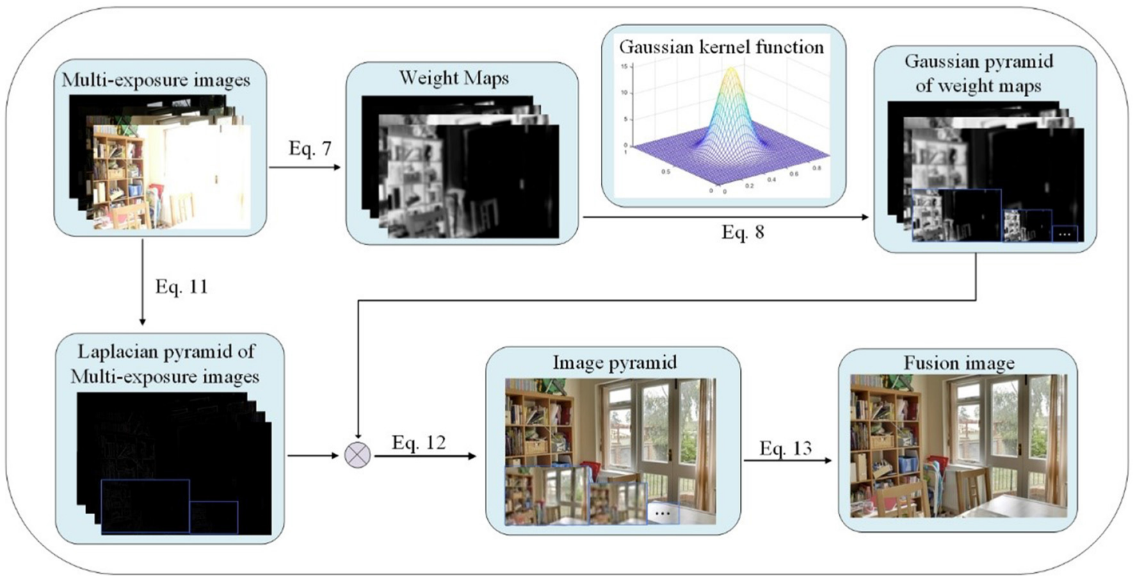


FIGURE 1 | The workflow of the proposed image fusion based on improved weight function.

pixel weight of the *n*th image. The workflow of the proposed image fusion based on improved weight function is shown in **Figure 1**. Equation (7), Equation (8) and other symbols in **Figure 1** correspond to the formula below, indicating that the operation corresponding to the equation has been performed. The symbol before Equation (12) in **Figure 1** represents the multiplication sign.

3. WEIGHT FUNCTION

As the core part of the proposed image fusion method, a reasonable weight function is designed based on the appropriate evaluation of exposure levels (Shen-yu et al., 2015). Gray value, as an important measure of image visible information, usually determines the fusion weight based on the distance between image gray and 0.5, but this single index will cause the loss of information of the fused image and some areas of the image are dark. Using the Evaluation of Moderate Exposure, the fusion weight is determined by the gray mean value of the multi exposure image at a certain point and the distance from 0.5 to retain more image information. Additionally, the relative brightness is applied to measure the corresponding weight.

3.1. Evaluation of Moderate Exposure

In the evaluation process, the brightness and darkness changes of different pixels obtained by the limited sampling of a scene are analyzed, and each image pixel value in the scene under the optimal moderate exposure is estimated. The differences between the pixel values of each input image and the corresponding optimal pixel values are compared to evaluate moderate exposure. The evaluation value can be directly used as the corresponding weight value for image fusion. For N images with different exposures from the same scene, $I_n(x, y)$ represents the pixel value at the coordinate (x, y) of the *n*th image, and the evaluation indicator of moderate exposure is the sum of weights used to obtain the fused image.

$$W_{1,n}(x,y) = \exp\{-\frac{(I_n(x,y) - \mu(x,y))^2}{2\delta^2}\} \tag{2}$$

$$\mu(x,y) = (1-\beta) * 0.5 + \beta * \bar{I}(x,y) \tag{3}$$

$$\bar{I}(x,y) = \frac{1}{N}\sum_{n=1}^{N} I_n(x,y) \tag{4}$$

In Equation (2), $\mu(x, y)$ represents the optimal pixel value of the pixel at the coordinate (x, y) of the image, which is estimated by Equation (3). On one hand, the value of $\mu(x, y)$ should be around 0.5, which can ensure ideal human visual experience. On the other hand, in order to reflect the real-world light-dark contrast information, it is necessary to approximate the brightness information from the limited sampling of the scene. Therefore, the average value of each pixel in the images with different exposures is calculated by Equation (4). $\mu(x, y)$ is the weighted sum of 0.5 and this average value. The weight factor β is a balance parameter between detail information and light-dark contrast information.

3.2. Relative Brightness

The evaluation indicator of moderate exposure cannot well capture the information from dark areas of long-exposure images or bright areas of short-exposure images. Therefore, the relative brightness proposed by Lee et al. (2018) is added as another weight indicator. Specifically, when the overall image is bright (long exposure), the relatively dark areas are given greater weights. Conversely, when the overall image is dark (short exposure), the relatively bright areas are given greater weights. The average pixel intensity of the nth image is denoted as m_n. When $I_n(x,y)$ is close to $1 - m_n$, the corresponding weight should be relatively large. Therefore, the relative brightness can be expressed as follows.

$$W_{2,n}(x,y) = \exp\{-\frac{(I_n(x,y) - (1 - m_n))^2}{2{\delta_n}^2}\} \tag{5}$$

In addition, when the adjacent exposed images and the input images have relatively large differences, the different objects in the two images are often in a good exposure state. Therefore, when the average brightness m_n of the nth image considerably differs from the average brightness m_{n-1},m_{n+1} of adjacent images, a larger δ_n value is given.α is a constant with a value of 0.75. δ_n controls the weight according to the different m_n values of the image, which can be expressed as follows.

$$\delta_n = \begin{cases} 2\alpha * (m_{n+1} - m_n), n = 1 \\ \alpha * (m_{n+1} - m_{n-1}), 1 < n < N \\ 2\alpha * (m_n - m_{n-1}), n = N \end{cases} \tag{6}$$

Therefore, the final weight function can be expressed as follows.

$$W_n(x,y) = W_{1,n}(x,y) * W_{2,n}(x,y) \tag{7}$$

4. MULTI-SCALE IMAGE DECOMPOSITION

Because the pixels of the image are closely related, it is more reliable to use a wider range of pixels to calculate the fusion weight. In addition, in the real world, objects have different structures at different scales. This shows that if you observe the same object from different scales, you will get different results. Therefore, in the case of multi-scale decomposition, using the image pyramid to calculate the result image will get better fusion results.

Gaussian pyramid decomposition is first performed on the weight map and the multi-exposure image sequences. Then the Laplacian pyramid decomposition is applied to the multi-exposure image sequences. After the Gaussian pyramid and Laplacian pyramid of the image are fused between the corresponding layers, the upper layer image of the fused pyramid is up-sampled, and the up-sampled image is added to the lower layer image to obtain an image with the equal size of the image to be fused.

4.1. Gaussian Pyramid Decomposition

The Gaussian pyramid obtains a series of down-sampled images through Gaussian smoothing and sub-sampling. Gaussian kernel is first used to convolve the image of the l layer, and then all even rows and columns are deleted to obtain the image of the $l + 1$ pyramid layer. The decomposition algorithm is shown as follows.

$$\begin{gathered} G_l(x,y) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m,n)G_{l-1}(2x+m, 2y+n) \\ (0 \le l \le L_{ev} - 1, 0 \le x \le C_l - 1, 0 \le y \le R_l - 1) \end{gathered} \tag{8}$$

where G_l is the image of the lth layer of the Gaussian Pyramid, C_l, R_l is the total number of rows and columns of the lth layer image, $w(m,n)$ is the value of the mth row and nth column of the Gaussian filter template, L_{ev} represents the number of Gaussian pyramid layers, and the maximum decomposable number of layers is $\log_2[\min(C_0, R_0)]$.

4.2. Laplace Pyramid Decomposition

The Gaussian pyramid obtained by Gaussian convolution and downsampling often loses detailed image information. Therefore, Mertens et al. (2010) introduced Laplacian pyramid to restore detailed image information. The image of each layer of Gaussian pyramid subtracts the predicted image obtained after the upsampling and Gaussian convolution of the upper layer image to obtain a series of difference images, which are the Laplacian decomposition images. First, the upsampling process is expressed as follows

$$expand(\widehat{G}_l(x,y)) = 4\sum_{m=-2}^{2}\sum_{n=-2}^{2} \widehat{G}_l(\frac{(x+m)}{2}, \frac{(y+n)}{2})w(m,n) \tag{9}$$

$$\widehat{G}_l(\frac{(x+m)}{2}, \frac{(y+n)}{2}) = \begin{cases} G_l(\frac{(x+m)}{2}, \frac{(y+n)}{2}), \frac{(x+m)}{2}, \frac{(y+n)}{2}) \in z \\ 0, else \end{cases} \tag{10}$$

where Z represents an integer, $expand(\widehat{G}_l(x,y))$ indicates that an upsampling operation is performed on the lth layer of Gaussian pyramid. As shown in Equation (11), the image G_l of the lth layer of Gaussian pyramid subtracts $expand(\widehat{G}_l(x,y))$ to obtain the lth layer image L_l containing detailed information.

$$L_l = \begin{cases} G_l - expand(\widehat{G}_l(x,y)), 0 \le l \le L_{ev} - 1 \\ G_{L_{ev}}, l = L_{ev} \end{cases} \tag{11}$$

The Laplace decomposition process of the image is shown in **Figure 2**. In this article, the number of layers of image pyramid is 7.

4.3. Image Fusion and Reconstruction

According to the above process, the Gaussian pyramid of the weighted image and the Laplacian pyramid of multi-exposure image sequences are first obtained, and then fused between the corresponding layers.

$$FI_l = \sum_{k=1}^{N} W_{k,l}L_{k,l}, 0 \le l \le L_{ev} - 1 \tag{12}$$

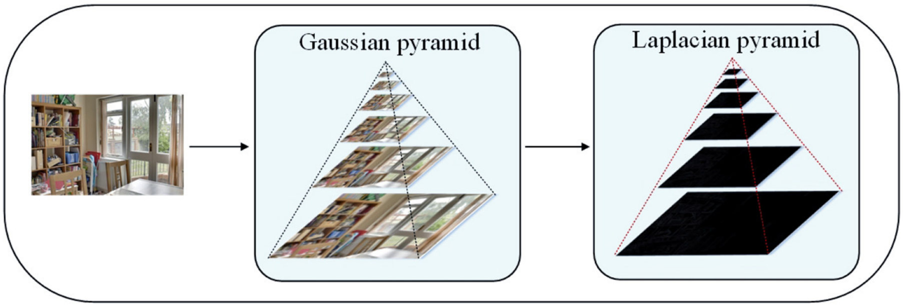


FIGURE 2 | The Laplace decomposition process of the image.

FI_l represents the fused image data of the lth layer. $W_{k,l}$ represents the lth layer data of the kth weighted image. $L_{k,l}$ represents the lth layer data of the Laplacian pyramid of the kth multi-exposure image. L_{ev} represents the total number of pyramid layers. N represents the number of images. The upper layer image of the fused pyramid is first upsampled, and then expanded and added to the lower layer image to obtain an image with the equal size of the image to be fused as follows.

$$H = \sum_{l=L_{ev}-2}^{0} FI_l + up(FI_{l+1}) \tag{13}$$

where FI_l represents the lth layer image of the fused pyramid, up represents upsampling, L_{ev} represents the number of pyramid levels, and H represents the final fusion image. The overall workflow of the proposed method is shown in **Algorithm 1**.

Algorithm 1 | Multi-exposure image fusion algorithm based on improved weight function.

```
Input LDR image sequences $I_k$ $k = 1, 2, \ldots N$, $N$ is the total number of images, $l$ represents the number of decomposition layers, $(x, y)$is the pixel position
Output the fused image
Calculation of image fusion weights:
for each $k \in [1, N]$ do
    $W_{(1,k)}(x, y) = exp(-(I_k(x, y) - \mu_k(x, y))^2/2\delta^2)$
    $W_{(2,k)}(x, y) = exp(-(I_k(x, y) - (1-\mathrm{m}_k))^2/2\delta_n^2)$
    $W_k(x, y) = W_{(1,k)}(x, y)W_{(2,k)}(x, y)$
end for
Gaussian pyramid decomposition of source image sequences and weight map:
for each $k \in [1, N]$ do
    for each $l \in [0, L_{ev} - 1]$ do
        $W_{k,l}(x, y) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m, n)W_{k,l}(2x + m, 2y + n)$
        $G_{k,l}(x, y) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m, n)I_{k,l-1}(2x + m, 2y + n)$
    end for
end for
Laplace Pyramid decomposition of source image sequences:
for each $k \in [1, N]$ do
    for each $l \in [0, L_{ev} - 1]$ do
        $\widehat{G}_l((x + m)/2, (y + n)/2) = \begin{cases} G_l(\frac{(x+m)}{2}, \frac{(y+n)}{2}), \frac{(x+m)}{2}, \frac{(y+n)}{2} \in z \\ 0, else \end{cases}$
        $expand(\widehat{G}_l(x, y)) = 4 \sum_{m=-2}^{2} \sum_{n=-2}^{2} \widehat{G}_l((x + m)/2, (y + n)/2)w(m, n)$
        $L_l = \begin{cases} G_l - expand(\widehat{G}_l(x, y)), 0 \le l \le L_{ev} - 1 \\ G_{L_{ev}}, l = L_{ev} \end{cases}$
    end for
end for
Image fusion reconstruction:
for each $k \in [1, N]$ do
    $FI_l = \sum_{k=1}^{N} W_{k,l}L_{k,l}, 0 \le l \le L_{ev} - 1$
end for
The fused image: $H = \sum_{l=L_{ev}-2}^{0} FI_l + up(FI_{l+1})$
```

5. EXPERIMENTAL RESULTS AND ANALYSIS

20 sets of multi-exposure image sequences, involving *Arno*, *Balloons*, *Cave*, *ChineseGarden*, etc., are applied to comparative experiments. The proposed method is subjectively and objectively compared with six existing MEF methods, including MESPD (Li et al., 2021), GD-MEF (Zhang and Cham, 2012), Fmmr (Li and Kang, 2012), DSIFT (Liu and Wang, 2015), GFF (Li et al., 2013), SPD-MEF (Ma et al., 2017) and PMEF (Qi et al., 2020). All experiments were performed in the matlab2019 environment on an Intel I7 9750H@2.60Ghz laptop with 8.00GB RAM. The relevant parameters are set to $\delta = 0.2$, $\beta = 0.5$ and $\alpha = 0.75$.

5.1. Subjective Comparison

Firstly, experiments are carried out on the "*Arno*" scene, and the fusion results of different algorithms are shown in **Figure 3**. It is not difficult to see that when dsift processes the clouds in the right sky, it is generally dark and can not capture the details of the clouds well. The GFF and SPD algorithms, when dealing with the bridge, have the problems of low brightness, resulting in the loss of detail information and poor visual effect. GD, PMEF and the algorithm proposed in this article can maintain the uniformity of the overall brightness of the image while retaining more details, and the visual effect is excellent.

The experimental results of "*Balloons*" scene are shown in **Figure 4**. The fusion results of GD, Fmmr and PMEF are dark. The details of clouds at the sunset are well captured, which

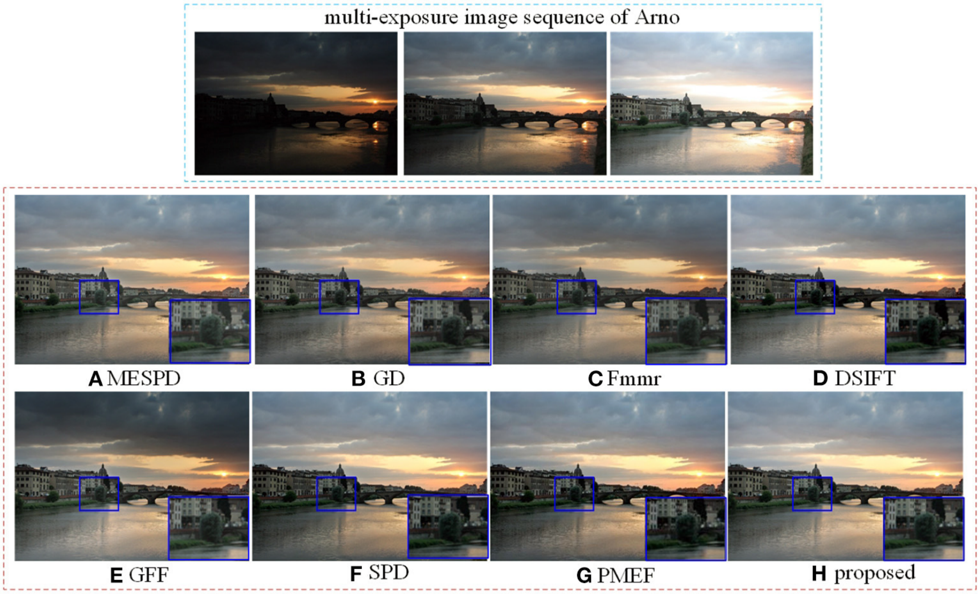


FIGURE 3 | Comparison of Arno scene experiment results of different methods.

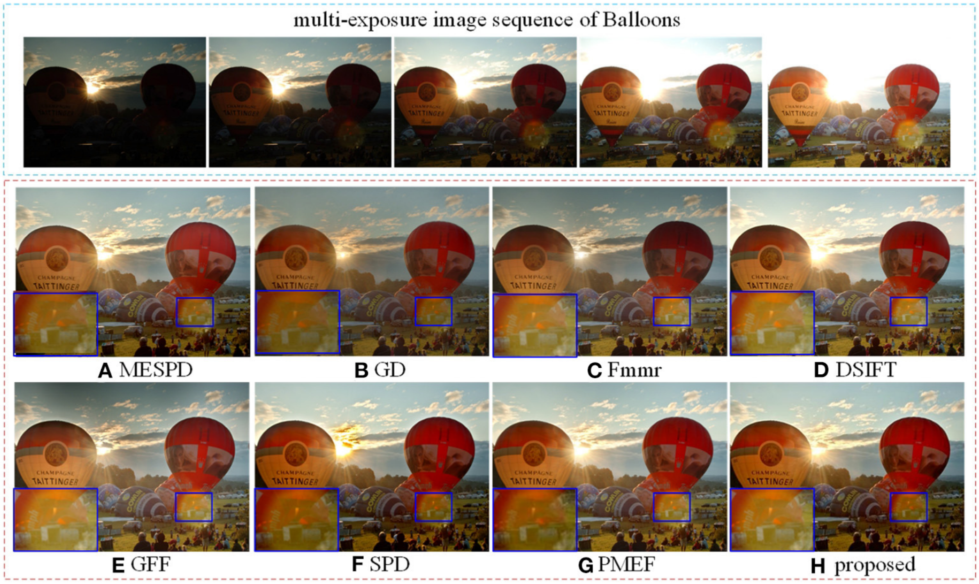


FIGURE 4 | Comparison of Ballons scene experiment results of different methods.

means the details of the overexposed image areas can be well captured. But the overall scene is dark, resulting in the detail loss of underexposed image areas. The image fused by GFF has a slight halo on the edge of the hot air balloon. Additionally, part of the sky is dark and the image color is slightly distorted. The sunset area of the image fused by SPD is abnormal. In addition, the image color is seriously distorted, which seriously affects the overall performance of the fused image. When comparing the enlarged details, MESPD, GD, Fmmr, SPD, and PMEF have low brightness, poor visibility and serious loss of details in this area.

In the experimental results of the "*Kluki*" scene, as shown in **Figure 5**, the saturation of SPD and PMEF is too high, resulting in some distortion of the color of the resulting image, and poor retention of the details of the clouds in the sky; Other algorithms retain the details of the clouds, and the visual effect is good. In contrast, the fusion results obtained by the proposed method and DSIFT consider the details of the bright and dark areas of the scene. So, the corresponding colors are real, the contrast is clear, and the visual performance of the fused images is good. From the enlarged details of the trees

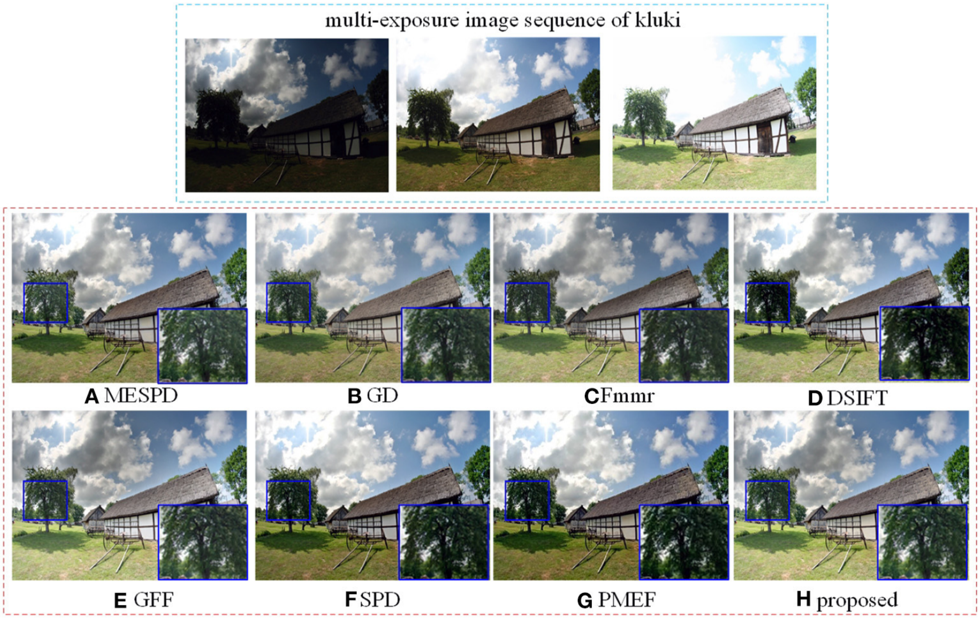


FIGURE 5 | Comparison of kluki scene experiment results of different methods.

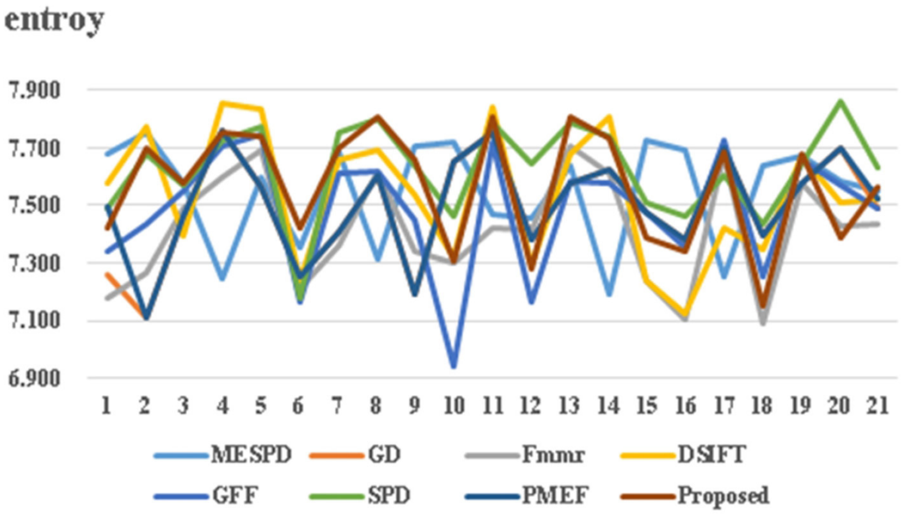


FIGURE 6 | Information entropy comparison of seven fusion methods.

on the left, dsift, SPD, and PMEF have the problems of low brightness and high saturation, resulting in poor retention effect of details.

5.2. Objective Evaluation Indicator Analysis

This article uses both structural similarity index (SSIM) and image information entropy for objective evaluation. As shown in **Figure 6** and **Tables 1**, **2**, the results confirm that the propose method achieves good performance in both subjective and objective evaluations. The abscissa in **Figure 6** represents the value of information entropy, and the abscissa in **Figure 7** represents the value of structural similarity; In addition, the ordinates of the two figures are the same: 1-20 represents different multi exposure sequences, and 21 represents the average value.

1) Image information entropy indicator comparison

Image information entropy is one of the important factors that determine the final effect of image fusion. The larger the information entropy, the more detailed information contained in the experimental result graph; On the contrary, the smaller the

TABLE 1 | Information entropy comparison of seven fusion methods.

Methods	MESPD	GD	Fmmr	DSIFT	GFF	SPD	PMEF	Proposed
1. Arno	**7.679**	7.258	7.175	7.581	7.343	7.490	7.498	7.424
2. Balloons	7.752	7.113	7.264	**7.773**	7.435	7.676	7.113	7.703
3. Cave	7.577	7.463	7.488	7.396	7.551	7.572	7.463	**7.579**
4. ChineseGarden	7.248	7.762	7.598	**7.852**	7.704	7.728	7.762	7.752
5. Church	7.601	7.565	7.693	**7.838**	7.744	7.774	7.565	7.737
6. Farmhouse	7.356	7.251	7.214	7.237	7.162	7.176	7.251	**7.424**
7. House	7.687	7.408	7.360	7.659	7.609	**7.755**	7.408	7.697
8. Kluki	7.312	7.603	7.620	7.696	7.618	7.801	7.603	**7.809**
9. Lamp	**7.705**	7.195	7.343	7.535	7.452	7.642	7.195	7.657
10. Landscape	**7.720**	7.655	7.303	7.322	6.938	7.460	7.655	7.305
11. Laurenziana	7.469	7.751	7.423	**7.840**	7.717	7.786	7.751	7.805
12. Lighthouse	7.458	7.384	7.413	7.292	7.167	**7.645**	7.384	7.283
13. MadisonCapitol	7.637	7.576	7.705	7.678	7.586	7.787	7.576	**7.809**
14. Mask	7.190	7.623	7.610	**7.811**	7.580	7.738	7.623	7.735
15. Office	**7.728**	7.473	7.236	7.236	7.473	7.507	7.473	7.387
16. Ostrow	**7.694**	7.382	7.105	7.122	7.356	7.460	7.382	7.342
17. Room	7.254	7.701	7.681	7.424	**7.729**	7.608	7.701	7.687
18. Set	**7.639**	7.394	7.092	7.347	7.255	7.438	7.394	7.150
19. Tower	7.672	7.576	7.579	7.675	7.657	7.646	7.576	**7.676**
20. Venice	7.584	7.701	7.430	7.513	7.571	**7.861**	7.701	7.387
21. Average	7.548	7.492	7.438	7.526	7.493	**7.633**	7.526	7.567

The bold value indicates the highest objective evaluation index value in this group of experiments.

TABLE 2 | Comparison of MEF-SSIM indexes of seven fusion methods.

Methods	MESPD	GD	fmmr	DSIFT	GFF	SPD	PMEF	Proposed
1. Arno	0.975	0.958	0.965	**0.989**	0.969	0.980	0.98	0.987
2. Balloons	0.959	0.893	0.945	0.968	0.948	0.965	0.965	**0.970**
3. Cave	**0.984**	0.964	0.961	0.972	0.978	0.948	0.969	0.980
4. ChineseGarden	0.987	0.982	0.982	**0.993**	0.984	0.985	0.986	0.989
5. Church	0.985	0.978	0.979	0.991	0.992	**0.993**	0.986	0.991
6. Farmhouse	0.970	0.971	0.977	0.976	0.985	**0.984**	0.977	0.978
7. House	**0.972**	0.865	0.926	0.964	0.957	0.898	0.941	0.953
8. Kluki	0.967	0.952	0.965	**0.981**	0.968	0.971	0.965	0.970
9. Lamp	0.968	0.972	0.972	0.973	0.942	**0.993**	0.983	0.965
10. Landscape	0.984	0.851	0.931	0.960	0.929	0.954	0.955	**0.983**
11. Laurenziana	0.98	0.982	0.976	0.989	0.987	**0.990**	0.982	0.986
12. Lighthouse	**0.979**	0.964	0.953	0.965	0.950	0.970	0.968	0.975
13. MadisonCapitol	**0.980**	0.932	0.918	0.973	0.968	0.977	0.973	**0.980**
14. Mask	0.987	0.975	0.982	**0.992**	0.979	0.988	0.981	0.990
15. Office	0.896	0.968	0.957	0.971	0.967	0.967	0.973	**0.988**
16. Ostrow	0.965	0.967	0.973	0.974	**0.986**	0.978	0.972	0.976
17. Room	0.976	0.975	0.973	**0.990**	0.960	0.988	0.984	0.980
18. Set	0.983	0.922	0.924	0.954	0.943	0.934	0.947	**0.984**
19. Tower	0.980	0.954	0.952	0.972	0.954	0.940	0.935	**0.985**
20. Venice	0.975	0.962	0.966	0.981	0.971	**0.982**	0.975	0.969
21. Average	0.973	0.949	0.959	0.976	0.966	0.969	0.970	**0.979**

The bold value indicates the highest objective evaluation index value in this group of experiments.

information entropy, the less detailed information contained in the experimental result graph. The evaluation results are shown in **Figure 6** and **Table 1**. The multi exposure fusion algorithm under multi-scale decomposition is slightly lower than the SPD algorithm based on image block decomposition and better than other algorithms. This is because the SPD algorithm based on image block decomposition avoids the partial loss of information caused by up and down sampling in multi-scale decomposition, and its entropy is better than the multi exposure fusion algorithm under multi-scale decomposition. The calculation formula of image entropy is as follows.

$$H = \sum_{i=0}^{255} P_i \log p_i \tag{14}$$

P_i represents the proportion of pixels with gray value i in the image.

2) MEF-SSIM comparison

This article uses the MEF quality evaluation model (Ma et al., 2015b) for evaluation. The proposed method is objectively compared with six existing MEF method. Natural images usually contain object information of different scales. Multi-scale can ensure the correlation between the pixels of different scales and optimize image fusion. Structural similarity as an index is used to measure the similarity of two images. As shown in **Figure 7** and **Table 2**, the MEF method under multi-scale decomposition achieves the best SSIM.

From the perspective of image composition, structural information is defined as an attribute that reflects the structure of objects in the scene independent of brightness and contrast. Additionally, model distortion is treated as a combination of three different factors, brightness, contrast, and structure. The mean is used as an estimate of brightness. The standard deviation is used as an estimate of contrast. The covariance is used as a measure of structural similarity. All the definitions are shown as follows.

$$SSIM(x, y) = [L(x, y)]^{\alpha} \cdot [C(x, y)]^{\beta} \cdot [S(x, y)]^{\gamma} \tag{15}$$

$$L(x, y) = \frac{2\mu_x\mu_y + c1}{\mu_x^2 + \mu_y^2 + c1} \tag{16}$$

$$C(x, y) = \frac{2\delta_x\delta_y + c2}{\delta_x^2 + \delta_y^2 + c2} \tag{17}$$

$$S(x, y) = \frac{\delta_{xy} + c3}{\delta_x\delta_y + c3} \tag{18}$$

$L(x, y)$, $C(x, y)$, and $S(x, y)$ are the comparison results of image brightness, contrast, and structure, respectively. μ_x and μ_y are the mean values of image pixels. δ_x and δ_y are the standard deviations of image pixel values. $\delta_{x,y}$ is the covariance of x and y. $c1$, $c2$, and $c3$ are constants to avoid system errors when the denominator is 0. α, β, γ used to adjust the weight of each component, usually α=β=γ=1. The structural similarity index is used for different scales, and the final image quality score is obtained through Formula (19).

$$MEF - SSIM = \sum_{l=1}^{L} [SSIM_l]^{\beta_l} \tag{19}$$

where L is the total number of scales and β_l is the weight assigned to the lth scale.

TABLE 3 | Ablation experiment of weight function.

Weight function	Evaluation of moderate	Exposure relative brightness	Proposed
Entroy	7.513	7.525	**7.567**
MEF-SSIM	0.966	0.970	**0.979**

The bold value indicates the highest objective evaluation index value in this group of experiments.

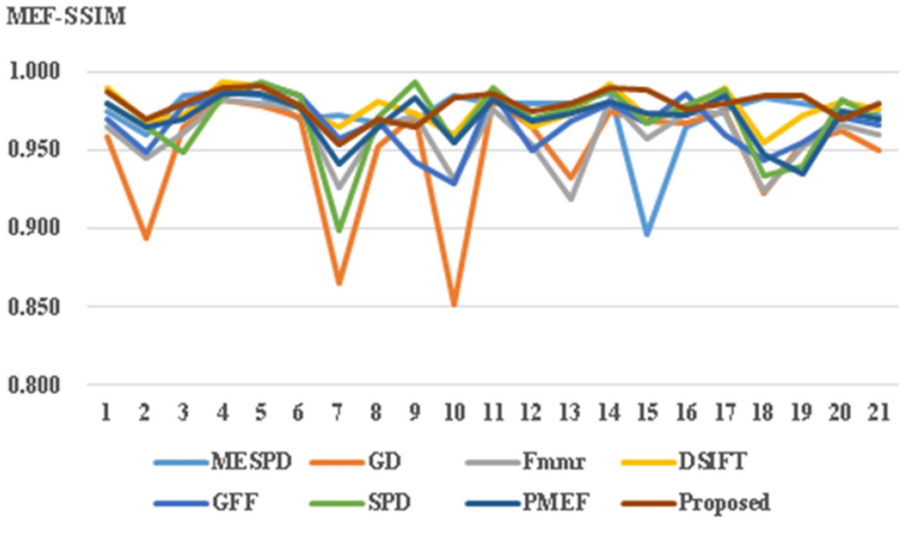


FIGURE 7 | Comparison of MEF-SSIM indexes of seven fusion methods.

TABLE 4 | Comparison of image fusion efficiency of seven fusion methods.

Methods	MESPD	GD	fmmr	DSIFT	GFF	SPD	PMEF	Proposed
Average time(s)	4.974	0.983	1.090	1.569	0.691	1.361	2.814	**0.327**

The bold value indicates the highest objective evaluation index value in this group of experiments.

5.3. Ablation Experiment of Weight Function

In order to prove that the weight function of two different feature indexes, moderate exposure evaluation and relative brightness, can make the multi exposure image fusion get better results. The following ablation experiments were carried out in this article. As shown in **Table 3**, the objective evaluation index of the fused image obtained by the improved weight function in this article performs well.

5.4. Comparison and Analysis of Computational Efficiency

As shown in **Table 4**, The computational efficiency of the multi exposure fusion algorithm based on the improved weight function is better than the comparison algorithm. In the multi-exposure fusion algorithm based on the improved weight function, although the Laplace image pyramid is used, in the continuous down sampling, the amount of calculation increases only a little due to the doubling of the number of pixels. In addition, because the weight calculation method of this algorithm is simple and easy to calculate, it does not need additional time. Therefore, the computational efficiency of this algorithm is obviously better than other comparison algorithms.

6. CONCLUSION

In this article, the weight function is improved, and the weight map is calculated by using the evaluation of moderate exposure and relative brightness. Pyramid-based multi-scale decomposition is used to fuse images with different resolutions to generate the final HDR image. The proposed method can effectively capture the rich image details and solve the issues such as splicing traces and border discontinuities in the fused image, avoiding the generation of artifacts. Both MEF-SSIM and image information entropy are used to evaluate the performance of image fusion. Experimental results confirm that the proposed method achieves good image fusion performance.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

KX: conceptualization, funding acquisition, and supervision. QW: investigation and methodology. HX and KL: software. All authors have read and agreed to the published version of the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Fei, K., Zhe, W., Chen, W., Wu, X., and Li, Z. (2017). Intelligent detail enhancement for exposure fusion. *IEEE Trans. Multimedia* 20, 484–495. doi: 10.1109/TMM.2017.2743988

Hayat, N., and Imran, M. (2019). Ghost-free multi exposure image fusion technique using dense sift descriptor and guided filter. *J. Vis. Commun. Image Represent.* 62, 295–308. doi: 10.1016/J.JVCIR.2019.06.002

Jin, X., Huang, S., Jiang, Q., Lee, S. J., and Yao, S. (2021a). Semi-supervised remote sensing image fusion using multi-scale conditional generative adversarial network with siamese structure. *IEEE J. Sel. Top. Appl. Earth Observations Rem. Sens.* doi: 10.1109/JSTARS.2021.3090958

Jin, X., Zhou, D., Jiang, Q., Chu, X., and Zhou, W. (2021b). How to analyze the neurodynamic characteristics of pulse-coupled neural networks? a theoretical analysis and case study of intersecting cortical model. *IEEE Trans. Cybern.* doi: 10.1109/TCYB.2020.3043233

Lee, S.-H., Park, J. S., and Cho, N. I. (2018). "A multi-exposure image fusion based on the adaptive weights reflecting the relative pixel intensity and global gradient," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens), 1737–1741.

Li, H., Chan, T. N., Qi, X., and Xie, W. (2021). Detail-preserving multi-exposure fusion with edge-preserving structural patch decomposition. *IEEE Trans. Circuits Syst. Video Technol.* doi: 10.1109/TCSVT.2021.3053405

Li, H., Ma, K., Yong, H., and Zhang, L. (2020). Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Trans. Image Process.* 29, 5805–5816. doi: 10.1109/TIP.2020.2987133

Li, S., and Kang, X. (2012). Fast multi-exposure image fusion with median filter and recursive filter. *IEEE Trans. Consum. Electron.* 58, 626–632. doi: 10.1109/TCE.2012.6227469

Li, S., Kang, X., and Hu, J. (2013). Image fusion with guided filtering. *IEEE Trans. Image Process.* 22, 2864–2875. doi: 10.1109/TIP.2013.2244222

Liu, C., Yuen, J., and Torralba, A. (2016). *SIFT Flow: Dense Correspondence Across Scenes and Its Applications*. Springer International Publishing, 15–49. doi: 10.1007/978-3-319-23048-1_2

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020). Multi-focus image fusion: a survey of the state of the art. *Inf. Fusion* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Liu, Y., and Wang, Z. (2015). Dense sift for ghost-free multi-exposure fusion. *J. Vis. Commun. Image Represent.* 31, 208–224. doi: 10.1016/j.jvcir.2015.06.021

Ma, K., Hui, L., Yong, H., Zhou, W., Meng, D., and Lei, Z. (2017). Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Trans. Image Process.* 26, 2519–2532. doi: 10.1109/TIP.2017.2671921

Ma, K., Yeganeh, H., Zeng, K., and Wang, Z. (2015a). High dynamic range image compression by optimizing tone mapped image quality index. *IEEE Trans. Image Process.* 24, 3086–3097. doi: 10.1109/TIP.2015.2436340

Ma, K., Zeng, K., and Wang, Z. (2015b). Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process. Publicat. IEEE Signal Process. Soc.* 24, 3345. doi: 10.1109/TIP.2015.2442920

Mertens, T., Kautz, J., and Reeth, F. V. (2010). Exposure fusion: a simple and practical alternative to high dynamic range photography. *Comput. Graph. Forum* 28, 161–171. doi: 10.1111/j.1467-8659.2008.01171.x

Moriyama, D., Ueda, Y., Misawa, H., Suetake, N., and Uchino, E. (2019). "Saturation-based multi-exposure image fusion employing local color correction," in *2019 IEEE International Conference on Image Processing (ICIP)* (Taipei), 3512–3516.

Qi, G., Chang, L., Luo, Y., Chen, Y., Zhu, Z., and Wang, S. (2020). A precise multi-exposure image fusion method based on low-level features. *Sensors (Basel, Switzerland)* 20, 1597. doi: 10.3390/s20061597

Shen, J., Zhao, Y., Yan, S., and Li, X. (2014). Exposure fusion using boosting laplacian pyramid. *IEEE Trans. Cybern.* 44, 1579–1590. doi: 10.1109/TCYB.2013.2290435

Shen, R., Cheng, I., Shi, J., and Basu, A. (2011). Generalized random walks for fusion of multi-exposure images. *IEEE Trans. Image Process.* 20, 3634–3646. doi: 10.1109/TIP.2011.2150235

Shen-yu, J., Kuo, C., Zhi-hai, X., Hua-jun, F., Qi, L., and Yue-ting, C. (2015). Multi-exposure image fusion based on well-exposedness assessment. *J. Zhejiang Univ. Eng. Sci.* (Hangzhou), 24, 7.

Ulucan, O., Karakaya, D., and Turkan, M. (2020). Multi-exposure image fusion based on linear embeddings and watershed masking. *Signal Process.* 178, 107791. doi: 10.1016/j.sigpro.2020.107791

Wang, S., and Zhao, Y. (2020). A novel patch-based multi-exposure image fusion using super-pixel segmentation. *IEEE Access* 8, 39034–39045. doi: 10.1109/ACCESS.2020.2975896

Xu, H., Jiang, G., Yu, M., Zhu, Z., Bai, Y., Song, Y., and Sun, H. (2021). Tensor product and tensor-singular value decomposition based multi-exposure fusion of images. *IEEE Trans. Multimedia* 1. doi: 10.1109/TMM.2021.3106789

Zhang, W., and Cham, W. K. (2012). Gradient-directed multiexposure composition. *IEEE Trans. Image Process.* 21, 2318–2323. doi: 10.1109/TIP.2011.2170079

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.

frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 10 June 2022
doi: 10.3389/fnbot.2022.823484

Cross Task Modality Alignment Network for Sketch Face Recognition

Yanan Guo [1,2], *Lin Cao* [1,2]* *and Kangning Du* [1,2]

[1] *Key Laboratory of Information and Communication Systems, Ministry of Information Industry, Beijing Information Science and Technology University, Beijing, China,* [2] *Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing, China*

The task of sketch face recognition refers to matching cross-modality facial images from sketch to photo, which is widely applied in the criminal investigation area. Existing works aim to bridge the cross-modality gap by inter-modality feature alignment approaches, however, the small sample problem has received much less attention, resulting in limited performance. In this paper, an effective Cross Task Modality Alignment Network (CTMAN) is proposed for sketch face recognition. To address the small sample problem, a meta learning training episode strategy is first introduced to mimic few-shot tasks. Based on the episode strategy, a two-stream network termed modality alignment embedding learning is used to capture more modality-specific and modality-sharable features, meanwhile, two cross task memory mechanisms are proposed to collect sufficient negative features to further improve the feature learning. Finally, a cross task modality alignment loss is proposed to capture modality-related information of cross task features for more effective training. Extensive experiments are conducted to validate the superiority of the CTMAN, which significantly outperforms state-of-the-art methods on the UoM-SGFSv2 set A, set B, CUFSF, and PRIP-VSGC dataset.

Keywords: sketch face recognition, cross-modality gap, small sample problem, image retrieval, feature alignment

OPEN ACCESS

Edited by:
Xin Jin,
Yunnan University, China

Reviewed by:
Aming Wu,
Tianjin University, China
Chuanbo Yu,
Tianjin University, China
Yuan Xue,
Beijing Jiaotong University, China

****Correspondence:***
Lin Cao
charlin@bistu.edu.cn

Received: *27 November 2021*
Accepted: *29 April 2022*
Published: *10 June 2022*

Citation:
Guo Y, Cao L and Du K (2022) Cross Task Modality Alignment Network for Sketch Face Recognition.
Front. Neurorobot. 16:823484.
doi: 10.3389/fnbot.2022.823484

1. INTRODUCTION

Face recognition plays an important role in law enforcement agencies (Lin et al., 2018). However, there are many cases where police cannot capture photos of a suspect, but eyewitnesses can help forensics draw a facial sketch. Sketch face recognition is the process of matching facial sketches to photos (Méndez-Vázquez et al., 2019); it has wide application in the criminal investigation area (Wang and Tang, 2009).

Sketch face recognition is challenging due to the large modality gap between photos and sketches and small sample problem. Photos depict the real-life environment. They have both macro edge and micro texture information. Sketches are usually hand-drawn (Wang and Tang, 2009) by forensic artists or composited (Galea and Farrugia, 2018) *via* computer software programs like EFIT-V and IdentiKit. They primarily contain macro edge information with minimal texture information. Moreover, due to the privacy protection problem and the time-consuming efforts of sketch drawing, amount of the paired sketch-photo data is limited, resulting in limited sketch face recognition performance. As a result, reducing the modality gap as much as possible has been important target in few shot sketch face recognition.

Several research studies have been devoted to reducing the modality gap, where it was divided into intra-modality (Gao et al., 2008b; Zhang et al., 2015) and inter-modality methods (Fan et al., 2020; Peng et al., 2021). For intra-modality methods, they aim to reduce the domain gap by transforming a sketch (photo) to a photo (sketch) first, and then using traditional homogeneous face recognition methods to match the resultant photos with the original photos. However, such methods usually contain undesirable artifacts (Zhang et al., 2015). Inter-modality methods aim to extract modality-invariant features to obtain promising performance. However, for small sample problem, these features usually are not optimal. Although several few-shot methods (Jiang et al., 2018; Dhillon et al., 2019) have achieved comparable performance on several benchmark datasets, they are not designed for sketch face recognition specifically and ignore an unavoidable fact that there exist modality shifts between sketch and photo domain.

In this paper, a Cross Task Modality Alignment Network (CTMAN) is proposed for sketch face recognition to address the above problem. Inspired by few-shot learning methods (Jiang et al., 2018), we introduced a meta learning training episode strategy to alleviate the small sample problem, several different tasks are built by the training episode strategy, then modality related query set and support set are designed to incorporate modality information. Based on these tasks, a two-stream network termed modality alignment embedding learning (MAE) is used to extract discriminative modality alignment features. Since mining important negative samples are important for few shot learning (Robinson et al., 2021), two cross task memory mechanisms are further proposed to obtain the cross task support set, thus the cross task support set can collect more sufficient hard negative features crossing different tasks (episodes), and the cross task modality alignment losses are computed over the cross task support set to enhance the discrimination of feature representations. Finally, by computing the distance between the query set and cross task support set, a cross task modality alignment loss is proposed to further guide the MAE to learn modality related features. Similar to Matching Networks (Xu et al., 2021) and Prototypical Networks (Snell et al., 2017), our proposed method can be seen as a form of meta-learning, in the sense that we compute the cross task domain alignment loss dynamically from new training tasks (episodes). The main difference between training episode strategy for few-shot learning and batch learning for traditional deep learning methods is that the label of identity in a different batch is fixed and in different episode is flexible.

Note that CTMAN is different from other sketch face recognition schemes, such as Domain Alignment Embedding Network (DAEN) (Guo et al., 2021). The main differences between the CTMAN and the DAEN are as follows: (1) CTMAN uses a two-stream network to extract discriminative modality alignment feature, the two-stream network consists of a ResNet50 backbone, the non-local blocks and the generalized mean (GeM) pooling layers. DAEN uses a traditional one-stream ResNet18 network to extract discriminative feature; (2) CTMAN proposes a cross task memory mechanism and cross task support feature set to collect more sufficient hard negative features by crossing different tasks and compute the cross task modality alignment losses over the query feature set and cross task support feature set. DAEN computes the modality alignment losses over the query feature set and support feature set.

Our major contributions can be summarized as follows: by utilizing the cross task information, we propose a CTMAN method to extract modality alignment discriminative representation under the small sample settings, achieving the competitive sketch face recognition performance. Furthermore, we design a cross task memory mechanism to obtain the updated cross task support set to collect more sufficient hard negative features by crossing different tasks. On the one hand, through manipulation of enqueue and dequeue, cross task memory mechanism can collect more sufficient hard negative features by crossing different tasks. On the other hand, by combining these hard negative features, the cross task support feature set is built for computing the cross task modality alignment losses to further enhance the discrimination of feature representations. The cross task modality alignment losses are computed over the query sketch feature set and cross task support feature set, they enhance feature representations by mining the modality relations between the sketch domain and photo domain. Extensive experimental results show that our proposed CTMAN outperforms the state-of-the-art methods on three benchmark datasets. Especially, on UoM-SGFSv2 set A and set B, our model achieves a significant improvement of 8.51 and 11.9% Rank-1, respectively, which greatly accelerates the sketch face recognition research.

The rest is arranged as follows. Previously related researches are briefly reviewed in Section 2. In Section 3, the CTMAN is introduced in detail. In Section 4, the experimental results on the UoM-SGFSv2 Set A, Set B, and CUFSF datasets are fully analyzed, and Section 5 concludes.

2. RELATED WORK

In this section, related sketch face recognition methods are reviewed. Since few-shot learning methods are related to our proposed method, these methods are also reviewed.

Sketch face recognition methods can be broadly divided into inter-modality and intra-modality methods. Eigen-transformation (Galea and Farrugia, 2015), Bayesian framework (Wang et al., 2017a), and Generative Adversarial Network (GAN) (Wang et al., 2017b) are representative intra-modality methods. Under the assumption that sketches and the corresponding photos are reasonably similar in appearance, the Eigen-transformation (Galea and Farrugia, 2015) used a linear combination of photos (or sketches) to synthesize whole images. Wang et al. (2017a) proposed a Bayesian framework to consider relationships among neighboring patch images for neighbor selection. With the development of GAN, many methods utilize GAN to transform a sketch into a photo. For example, Wan and Lee (2019) proposed a residual dense U-Net generator and a multitask discriminator for sketch face generation and recognition simultaneously. However, these methods do not emphasize inter-personal differences, causing performance reduction when data samples are limited, moreover, these methods are computationally expensive (Zhang et al., 2015).

Traditional inter-modality methods include the local binary pattern (LBP) (Bhatt et al., 2010), histogram of averaged orientation gradients (HAOG) (Galoogahi and Sim, 2012), and logGabor-MLBP-SROCC (LGMS) method (Galea and Farrugia, 2016). Bhatt et al. (2010) used extended uniform circular LBP descriptors to characterize sketches and photos. The HAOG (Galoogahi and Sim, 2012) is a gradient orientation based face descriptor, it was proposed to reduce the modality difference by the fact that gradient orientations of macro edge information are more modality invariant than micro texture information. By utilizing multiscale LBP and log-Gabor filters, Galea and Farrugia (2016) proposed LGMS method to extract local and global texture representations for sketch face recognition. Recently, many works attempt to address the cross-modal matching problem by deep learning methods benefiting from the development of deep learning (Mittal et al., 2015; Peng et al., 2019, 2021; Fan et al., 2020). Mittal et al. (2015) proposed a deep belief model to learn a feature of photos and then fine-tuned it for sketch face recognition. By introducing a soft face parsing approach, Peng et al. (2021) proposed a soft semantic representation method to extract contour level and soft semantic level deep features. They also proposed a deep local feature learning approach to learn compact and discriminant local information directly from original facial patches. Fan et al. (2020) presented a Siamese graph convolution network by building cross-modal graphs for face sketch recognition. However, the success of these deep learning approaches neglects the small sample problem to some extent.

By using a 3-D morphable model to synthesize both photos and sketches to augment the training data, Galea and Farrugia (2018) utilized a fine-tuned VGG-Face network and a triplet loss to determine the identity in a query sketch by comparing it to a gallery set. Guo et al. (2021) designed a training episode strategy to alleviate the small sample problem and proposed a domain alignment embedding loss to guide the network to learn discriminative features. Recently, few-shot learning has become appealing choice to deal with a small sample problem. Metric based meta-learning method and hard samples mining method are representative methods for few-shot learning. Metric based meta-learning method raises the learning level from data level to task level, and it learns the embedding from newly labeled tasks instead of the whole training dataset in each episode. Vinyals et al. (2016) proposed a matching network by using an attention mechanism to predict the class of query sets from labeled support sets. Wang J. et al. (2018) proposed a Siamese network by minimizing a pairwise similarity metric between within-class samples. By regarding each image as a graph node, Garcia and Bruna (2017) designed a Graph Neural Network to learn the information transmission task in an end-to-end manner. For the hard samples mining technique, Zhong et al. (2019) utilized the instance invariance technique in domain adaptation to construct positive exemplar memory. Wang et al. (2019) proposed a cross batch memory to provide a rich set of negative samples by using a dynamic queue of mini-batches. Robinson et al. (2021) developed an efficient and easy to implement sampling technique for selecting hard negative samples with few computational overheads. Although the above hard samples mining methods have achieved competitive performance on several representative small sample dataset, they do not consider the modality gap between sketch images and photo images.

3. PROPOSED METHOD

In this section, we detail the proposed CTMAN. Several training episodes are randomly selected from the training set to mimic few shot tasks, and modality related query set and support set are designed to incorporate domain information in meta learning training episode strategy stage. In each training episode, we use a MAE network to extract discriminative features to obtain the modality alignment query feature set and support feature set. On the basis of the support feature set, to further alleviate the small sample problem, we propose two cross task memory mechanism to obtain the cross task support set to collect sufficient hard negative features crossing different tasks. Finally, a cross task modality alignment loss is computed over the query feature set and cross task support feature set and a modality alignment loss is computed over the query feature set, and support feature set. **Figure 1** shows the proposed CTMAN in one training episode.

3.1. Meta Learning Episode Training Strategy

Due to the privacy protection problems and the time consuming efforts of sketch drawing, amount of the paired sketch-photo data is limited. Inspired by the few shot learning methods (Vinyals et al., 2016; Snell et al., 2017; Jiang et al., 2018; Guo et al., 2021), a meta learning training episode strategy is introduced to incorporate modality information by sampling image pairs and classes from the training set.

Given a training set $D_{tr} = \{S, P\} = \{(s_1, y_1), \cdots, (s_N, y_N), (p_1, y_1), \cdots, (p_N, y_N)\}$, where $P = \{(p_i, y_i)\}_{i=1}^{N}$ are photo images and $S = \{(s_i, y_i)\}_{i=1}^{N}$ are sketch images, N is the number of subjects, y_i is the class label, s_i and $p_i (i = 1 : N)$ share same label y_i. The meta learning training episode classes $B = \{t_1, \ldots, t_b\} \subset \{1, \cdots, N\}$ is randomly selected to form the meta learning training episode or task $D^t = \{(s_1^t, y_1^t, 1), \cdots, (s_b^t, y_b^t, b), (p_1^t, y_1^t, 1), \cdots, (p_b^t, y_b^t, b)\}$, where $s_k^t = s_{i_k}, p_k^t = p_{i_k}, y_k^t = y_{i_k}, k = 1, \cdots, b$, y_k^t is original label corresponding to s_k^t and p_k^t, and k is the current label corresponding to s_k^t and p_k^t in the current training episode. For each training epoch, the meta learning training episode D^t will be randomly formulated T times $(D^1, \cdots, D^T)$ to mimic the few-shot task.

In each training episode D^t, a query set $Q^t = \{(s_1^t, 1), \cdots, (s_b^t, b), (p_1^t, 1), \cdots, (p_b^t, b)\}$ is builded. For $s_i^t \in Q^t, i = 1, \cdots, b$, the corresponding photo support set is builded by $S_p^t = \{(p_1^t, y_1^t, 1), \cdots, (p_b^t, y_b^t, b)\}$. For $p_i^t \in Q^t$, the corresponding sketch support set is builded by $S_s^i = \{(s_1^t, y_1^t, 1), \cdots, (s_b^t, y_b^t, b)\}$.

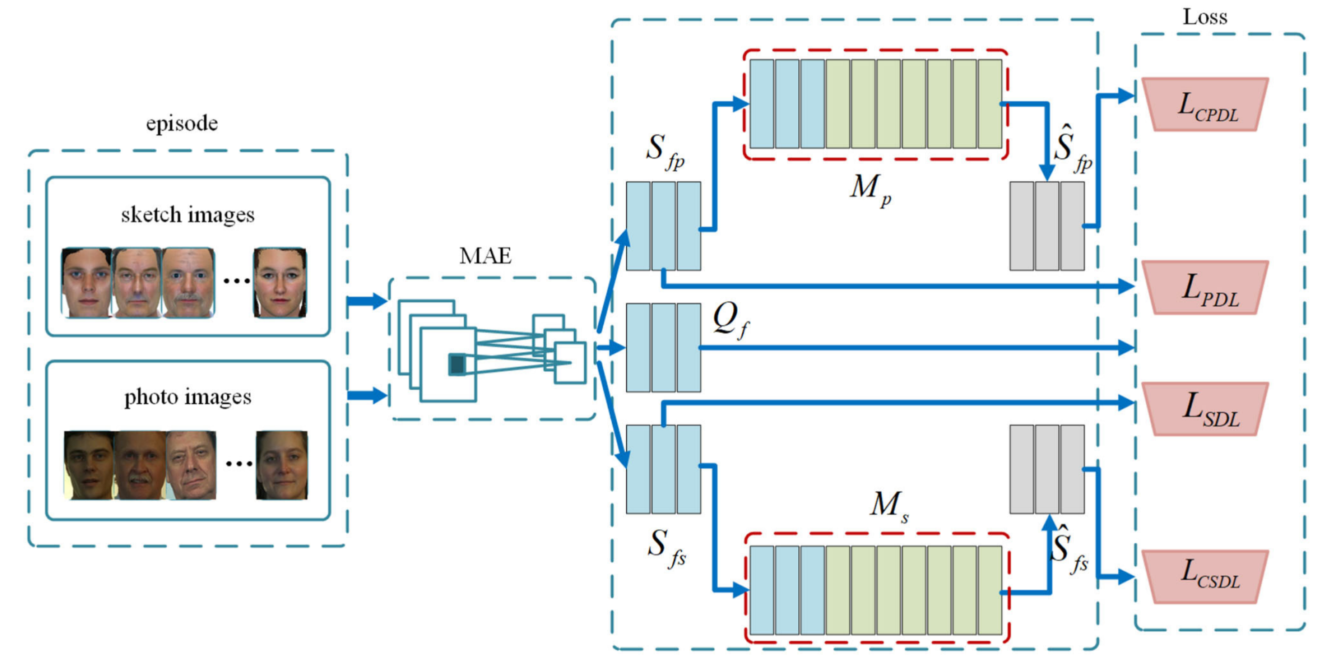


FIGURE 1 | Cross task modality alignment network (CTMAN) for sketch face recognition. In each training episode, CTMAN first embeds sketch and photo images into feature space by modality alignment embedding (MAE) network to obtain query feature set Q_f and support feature set (S_{fs} and S_{fp}). Then, it proposes two cross task memory mechanism M_p and M_s to obtain the cross task support feature set ($\hat{S}_{fs}$ and $\hat{S}_{fp}$). Finally, a cross task modality alignment loss ($L_{CPDL} + L_{CSDL}$) is computed on the query feature set and cross task support feature set, a modality alignment loss ($L_{PDL} + L_{SDL}$) is computed on the query feature set and the support feature set, and the final loss is computed over the cross task modality alignment loss and modality alignment loss.

3.2. Modality Alignment Embedding Learning

Since two-stream network structure has been widely used in cross-modality person re-identification and achieved comparable performance (Ye et al., 2020), here we introduce a two-stream feature extraction network structure (Ye et al., 2021) termed MAE network $F(\cdot) = [F_s(\cdot), F_p(\cdot)]$ for sketch face recognition to capture more modality-specific and modality-sharable features. The overall structure of MAE for sketch face recognition is illustrated in **Figure 2**. The structure of ResNet50 (He et al., 2016) pre-trained on ImageNet is adopted as a backbone for MAE, and the fully connected layer is removed. The MAE contains two blocks, the first block is designed specifically for two modalities in order to capture modality-specific information while the remaining blocks are shared to learn modality-sharable features. The first block contains a convolutional layer, a batchnorm layer, a relu layer, and a maxpooling layer. The remaining blocks contain 4 residual modules and 4 non-local attention blocks (Wang et al., 2017c), each residual module follows a non-local attention blocks, the final non-local attention block follows a pooling layer, the output of the pooling layer is adopted for computing loss function in the training and inference stage. Since sketch face recognition is a cross modal fine-grained instance retrieval, the widely-used max-pooling or average pooling cannot capture the domain-specific discriminative features (Ye et al., 2021), here we adopt a GeM pooling (Radenovic et al., 2017) for the pooling layer.

In each training episode D^t, a query set Q^t, a photo support set S^t_p, and sketch support set S^t_s are given. $F(\cdot) = [F_s(\cdot), F_p(\cdot)]$ embeds them to the query feature set $Q_f = \{(F_s(s^t_1), 1), \cdots, (F_s(s^t_b), b), (F_p(p^t_1), 1), \cdots, (F_p(p^t_b), b)\} = \{(f^t_{s1}, 1), \cdots, (f^t_{sb}, b), (f^t_{p1}, 1), \cdots, (f^t_{pb}, b)\}$, photo support feature set $S_{fp} = \{(F_p(p^t_1), y^t_1, 1), \cdots, (F_p(p^t_b), y^t_b, b)\} = \{(f^t_{p1}, y^t_1, 1), \cdots, (f^t_{pb}, y^t_b, b)\}$, and sketch support feature set $S_{fs} = \{(F_s(s^t_1), y^t_1, 1), \cdots, (F_s(s^t_b), y^t_b, b)\} = \{(f^t_{s1}, y^t_1, 1), \cdots, (f^t_{sb}, y^t_b, b)\}$, respectively.

3.3. Cross Task Modality Memory Mechanism

Mining important negative samples are important for few shot learning (Robinson et al., 2021) and metric learning (Wang et al., 2019), for collecting sufficient informative negative pairs from each episode, inspired by Wang et al. (2019), through the manipulation of enqueue and dequeue. We propose a cross task photo memory mechanism M_p and a cross task sketch memory mechanism M_s to record the deep features of recent episodes, allowing the model to collect sufficient hard negative pairs across multiple tasks. By computing the mean value of within class sample of the M_p and M_s, a cross task photo support feature set

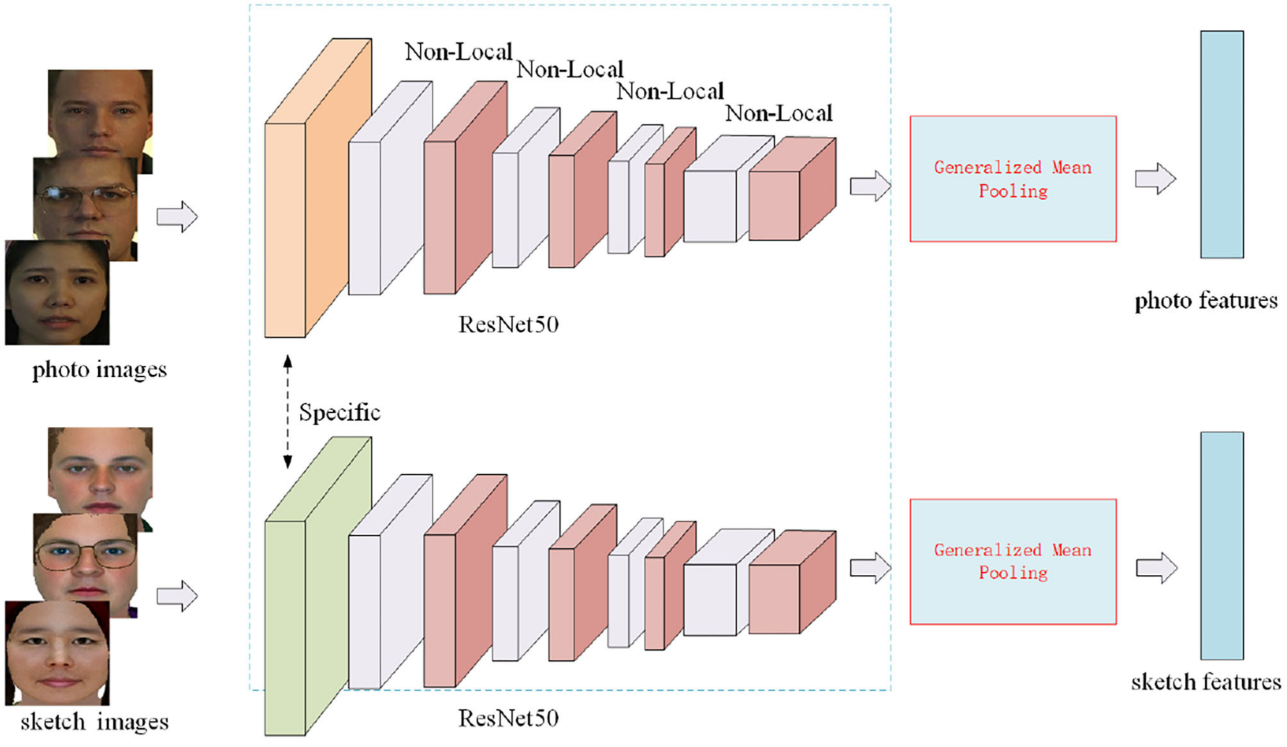


FIGURE 2 | The overall structure of MAE for sketch face recognition.

$\hat{S}_{fp}$ and a cross task sketch support feature set $\hat{S}_{fs}$ are obtained for computing the cross task modality alignment losses to enhance the discrimination of feature representations.

Suppose M is the memory size of M_p and $b < M$, the $M_p = \{(\bar{f}_{p1}, \bar{y}_1), \cdots, (\bar{f}_{pM}, \bar{y}_M)\}$ and $\hat{S}_{fp}$ are builded and updated as follows: in the first m episode, the MAE is warmed up first to reach a local optimal field, $M_p = \{(\bar{f}_{p1}, \bar{y}_1), \cdots, (\bar{f}_{pM}, \bar{y}_M) = \{(f^m_{p1}, y^m_1), \cdots, (f^m_{pb}, y^m_b), (0,0), \cdots, (0,0)\}$, $\hat{S}_{fp} = S_{fp} = \{(f^m_{p1}, y^m_1, 1), \cdots, (f^m_{pb}, y^m_b, b)\}$. Then, for the following task, the features and original labels of the current task of M_p are enqueued and entities of the earliest task are dequeued. For example, for the $(m+1)_{th}$ episode, if $2b \leq M$, the M_p is updated by $M_p = \{(f^m_{p1}, y^m_1), \cdots, (f^m_{pb}, y^m_b), (f^{m+1}_{p1}, y^{m+1}_1), \cdots, (f^{m+1}_{pb}, y^{m+1}_b), (0,0), \cdots, (0,0)\}$, else if $2b - M = k \geq 0$, $M_p = \{(f^m_{p(k+1)}, y^m_{k+1}), \cdots, (f^m_{pb}, y^m_b), (f^{m+1}_{p1}, y^{m+1}_1), \cdots, (f^{m+1}_{pb}, y^{m+1}_b)\}$. The $\hat{S}_{fp}$ is updated by $\hat{S}_{fp} = \{(\hat{f}^{m+1}_{p1}, y^{m+1}_1, 1), \cdots, (\hat{f}^{m+1}_{pb}, y^{m+1}_b, b)\}$, for each $\hat{f}^{m+1}_{pi}$ with label y^{m+1}_i, suppose there exist q_i with-in class feature in M_p selected by label y^{m+1}_i, then $\hat{f}^t_{pi}$ is computed by

$$\hat{f}^{m+1}_{pi} = \frac{1}{q_i + 1}\left(\sum_{\bar{y}_n = y^{m+1}_i, \bar{f}_{pn} \neq f^{m+1}_{pi}} \bar{f}_{pn} + f^{m+1}_{pi}\right). \tag{1}$$

Likewise, a cross task sketch memory mechanism $M_s = \{(\bar{f}_{s1}, \bar{y}_1), \cdots, (\bar{f}_{sM}, \bar{y}_M)\}$ and a cross task sketch support feature set $\hat{S}_{fs} = \{(\hat{f}^t_{s1}, y^t_1, 1), \cdots, (\hat{f}^t_{sb}, y^t_b, b)\}$ can be builded in a similar way, suppose there exist h_i with-in class feature in M_p selected by label y^t_i, $\hat{f}^t_{si}$ is computed by

$$\hat{f}^t_{si} = \frac{1}{h_i + 1}\left(\sum_{\bar{y}_n = y^t_i, \bar{f}_{sn} \neq f^t_{si}} \bar{f}_{sn} + f^t_{si}\right). \tag{2}$$

3.4. Cross Task Modality Alignment Loss

Based on the above meta learning training episode strategy and cross task modality memory mechanism, a cross task modality alignment loss is proposed and a modality alignment loss is used to guide the $F(\cdot)$ to learn discriminative modality alignment features. In each training episode, the query feature set $Q_f = \{(f^t_{s1}, 1), \cdots, (f^t_{sb}, b), (f^t_{p1}, 1), \cdots, (f^t_{pb}, b)\}$, photo support feature set $S_{fp} = \{f^t_{p1}, y^t_1, 1), \cdots, f^t_{pb}, y^t_b, b)$, and sketch support feature set $S_{fs} = \{f^t_{s1}, y^t_1, 1), \cdots, f^t_{sb}, y^t_b, b)$ are extracted by the MAE learning $F(\cdot)$ first. Then, the cross task photo support feature set $\hat{S}_{fp} = \{(\hat{f}^t_{p1}, y^t_1, 1), \cdots, (\hat{f}^t_{pb}, y^t_b, b)\}$ and cross task sketch support feature set $\hat{S}_{fs} = \{(\hat{f}^t_{s1}, y^t_1, 1), \cdots, (\hat{f}^t_{sb}, y^t_b, b)\}$ are builded by cross task modality memory mechanism.

For a sketch feature f^t_{si} in query feature set Q_f, its probability distribution over the cross task photo support set $\hat{S}_{fp}$ can be formulated by a softmax function over b cross task photo features:

$$P(k|f_{si}^t) = \frac{\exp(-||f_{si}^t - \hat{f}_{pk}^t||)}{\sum_{j=1}^{b} \exp(-||f_{si}^t - \hat{f}_{pj}^t||)}, \tag{3}$$

where $||\cdot||$ is the Frobenius norm, $P(k|f_{si}^t)$ refers to the probability of s_i^t belonging to the class k.

By summarizing the probability $P(k|f_{si}^t)$, $i = 1, \cdots, b$ on the Q_f, the cross task sketch modality embedding loss is denoted as follows:

$$L_{CSDL} = \frac{1}{b}\sum_{i=1}^{b} -\log P(k|f_{si}^t), \tag{4}$$

Similarly, the cross task photo modality embedding loss L_{CPDL} is denoted as follows:

$$L_{CPDL} = \frac{1}{b}\sum_{i=1}^{b} -\log P(k|f_{pi}^t) = \frac{1}{b}\sum_{i=1}^{b} -\log(\frac{\exp(-||f_{pi}^t - \hat{f}_{sk}^t||)}{\sum_{j=1}^{b} \exp(-||f_{pi}^t - \hat{f}_{sj}^t||)}), \tag{5}$$

Combine Equations (4) and (5), the cross task modality alignment loss is computed by the sum of the cross task sketch domain embedding loss and the cross task photo domain embedding loss:

$$\begin{aligned} L_{CDL} &= \frac{1}{2}(L_{CPDL} + L_{CSDL}) \\ &= \frac{1}{2b}(\sum_{i=1}^{b} -\log P(k|f_{pi}^t) + \sum_{i=1}^{b} -\log P(k|f_{si}^t)). \end{aligned} \tag{6}$$

To further extract discriminative modality alignment features, the probability distribution of Q_f over the photo support set S_{fp} and sketch support set S_{fs} are also computed as follows:

$$P_1(k|f_{si}^t) = \frac{\exp(-||f_{si}^t - f_{pk}^t||)}{\sum_{j=1}^{b} \exp(-||f_{si}^t - f_{pj}^t||)}, \tag{7}$$

$$P_1(k|f_{pi}^t) = \frac{\exp(-||f_{pi}^t - f_{sk}^t||)}{\sum_{j=1}^{b} \exp(-||f_{pi}^t - f_{sj}^t||)}, \tag{8}$$

Finally, the modality alignment loss is computed by the sum of the sketch domain embedding loss L_{PDL} and the photo domain embedding loss L_{SDL}:

$$L_{DL} = L_{PDL} + L_{SDL} = \frac{1}{2b}(\sum_{i=1}^{b} -\log P_1(k|f_{pi}^t) + \sum_{i=1}^{b} -\log P_1(k|f_{si}^t)), \tag{9}$$

Combine Equations (6) and (9), the final loss is computed by the weight sum of the cross task modality alignment loss and the modality alignment loss:

$$\begin{aligned} L &= \frac{1}{2}(L_{DL} + \lambda L_{CDL}) \\ &= \frac{1}{2b}(\sum_{i=1}^{b} -\log P_1(k|f_{pi}^t) + \sum_{i=1}^{b} -\log P_1(k|f_{si}^t)) \\ &+ \frac{\lambda}{2b}(\sum_{i=1}^{b} -\log P(k|f_{pi}^t) + \sum_{i=1}^{b} -\log P(k|f_{si}^t)). \end{aligned} \tag{10}$$

where λ is the trade-off parameter.

3.5. Learning and Inference

For each episode, we update the parameter of MAE by the solving following optimization problem:

$$\min_{w} L = \frac{1}{2}(L_{DL} + \lambda L_{CDL}). \tag{11}$$

The detailed process of loss computation is provided in Algorithm 1, which can be optimized with back-propagation algorithm. As for inference, after extracting the probe feature set and gallery feature set from the well-trained MAE network $F(\cdot) = [F_s(\cdot), F_p(\cdot)]$, for each sketch feature $F_s(s^e)$ in probe feature set, we compute Euclidean metric among the $F_s(s^e)$ and the gallery feature set $\{F_p(p^1), \cdots, F_p(p^n)\}$, the corresponding nearest gallery sample p_i^e is the matched photo image.

Algorithm 1: Loss computation of CTMAN.

```
Input: training episode $D^t = \{(s_1^t, y_1^t, 1), \cdots, (s_b^t, y_b^t, b),$
       $(p_1^t, y_1^t, 1), \cdots, (p_b^t, y_b^t, b)\}$.
Build a query set $Q^t$, a photo support set$S_p^t$, and a sketch
  support set $S_s^t$ by Section 3.1;
Build a query feature set $Q_f$, a photo support feature set$S_{fp}$,
  and a sketch support feature set $S_{fs}$ by Section 3.2;
Build a cross task photo support feature set $\hat{S}_{fp}$ and a cross
  task sketch support feature set $\hat{S}_{fs}$ by Section 3.2;
Compute the cross task modality alignment loss $L_{CDL}$ and
  modality alignment loss $L_{DL}$ by Equation (6) and
  Equation (9), respectively;
Compute $L$ by Equation (11) ;
Output: $L$.
```

4. EXPERIMENT

The proposed CTMAN is evaluated through extensive experiments on the UoM-SGFSv2 dataset (Galea and Farrugia, 2018) and the CUHK Face Sketch FERET Database (CUFSF) dataset (Mittal et al., 2015). Extensive ablation analysis is conducted to verify effectiveness of each contribution of the CTMAN. Finally, the proposed method is compared with other most recent competing methods on sketch face accuracy.

TABLE 1 | Experiment setup, UoM-SGFS set A* is UoM-SGFS set A, MEDS -II, FEI, and LFW, and UoM-SGFS set B* is UoM-SGFS set B, MEDS -II, FEI, and LFW.

Setup name	Training set	Test set	Train/pairs	Probe	Gallery
S1	UoM-SGFSv2 set A	UoM-SGFS set A*	450	150	150+1521
S2	UoM-SGFSv2 set B	UoM-SGFS set B*	450	150	150+1521
S3	CUFSF	CUFSF	500	694	694
S4	PRIP-VSGC	PRIP-VSGC	48	75	75

4.1. Dataset

The UoM-SGFSv2 database (Galea and Farrugia, 2018) consists of 600 paired sketch and photo samples. The 600 photos come from the Color-FERET database (Rallings et al., 1998), for each of the 600 photos, two viewed sketches were drawn by computer. One viewed sketch was drawn using EFIT-V software manually operated by an artist, and the other was further edited utilizing the Image editing software, thus, the other is closer in appearance to the photos. The UoM-SGFSv2 set A consists of 600 photos, and the 600 sketches is drawn using the EFIT-V software, and the UoM-SGFSv2 set B consists of the 600 photos and the other 600 sketches. The CUFSF dataset contains 1,194 subjects, each subject has one photo image with illumination changes coming from the FERET database (Rallings et al., 1998) and one sketch image created by an artist. This database is challenging due to the different illumination conditions of the photo images and several exaggerations of the sketch images. The PRIP-VSGC dataset contains 123 subjects, each subject has one photo that comes from the AR dataset (Martinez and Benavente, 1998), and one sketch created by an Asian artist by utilizing the Identi-Kit tool.

Based on the above three datasets, four experimental setup are performed. S1 setup and S2 setup are based on the UoM-SGFSv2 set A and B, respectively, and the partition protocols in Galea and Farrugia (2018) are followed. The training set consists of 450 randomly selected subjects, and the test set contains the rest 150 subjects. When tested, the 150 sketch images form the probe set and 150 photo images form the gallery set, to mimic the mug-shot galleries, the gallery set is further extended to 1,521 subjects. These 1,521 subjects include 199 subjects from the FEI dataset[1],

[1] Available at: http://fei.edu.br/~cet/facedatabase.html.

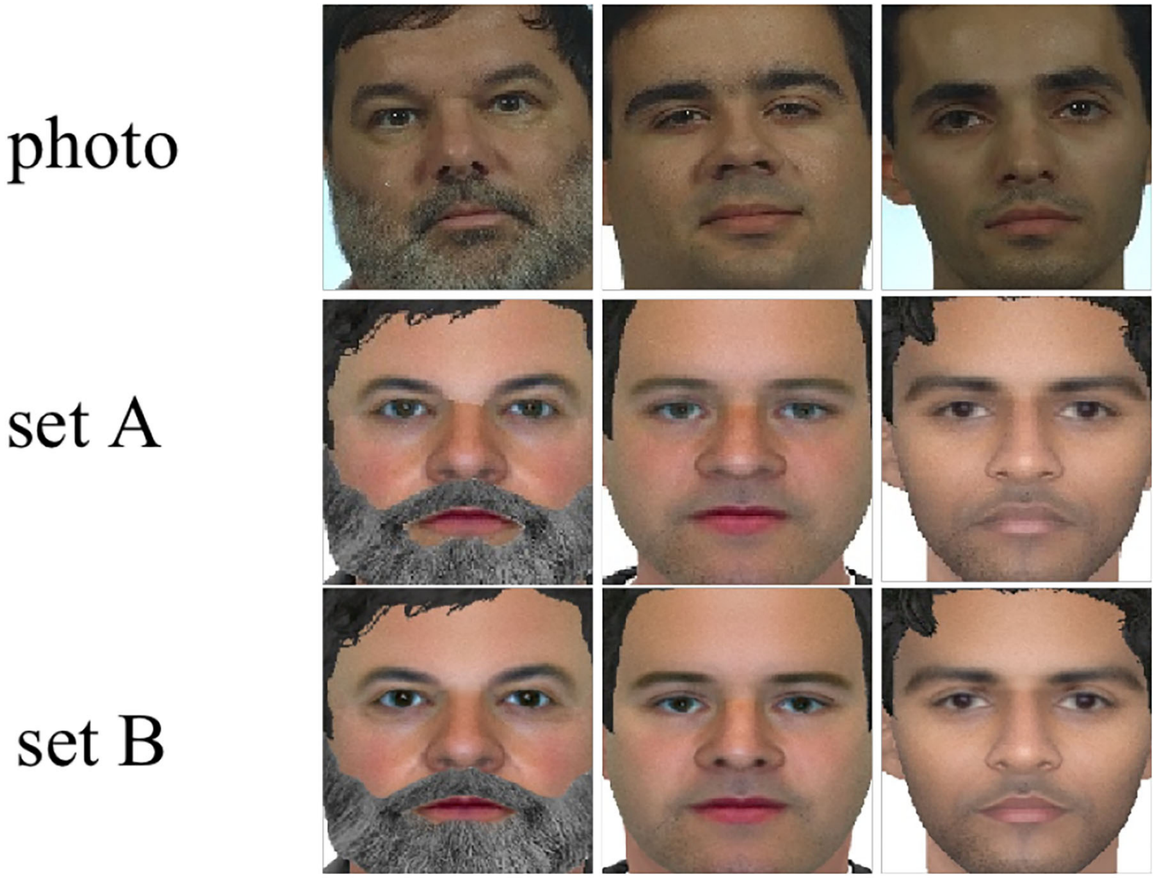


FIGURE 3 | Examples of cropped images from the UoM-SGFSv2 dataset, the top, middle, and bottom row are photo images, sketch images from set A and set B, respectively.

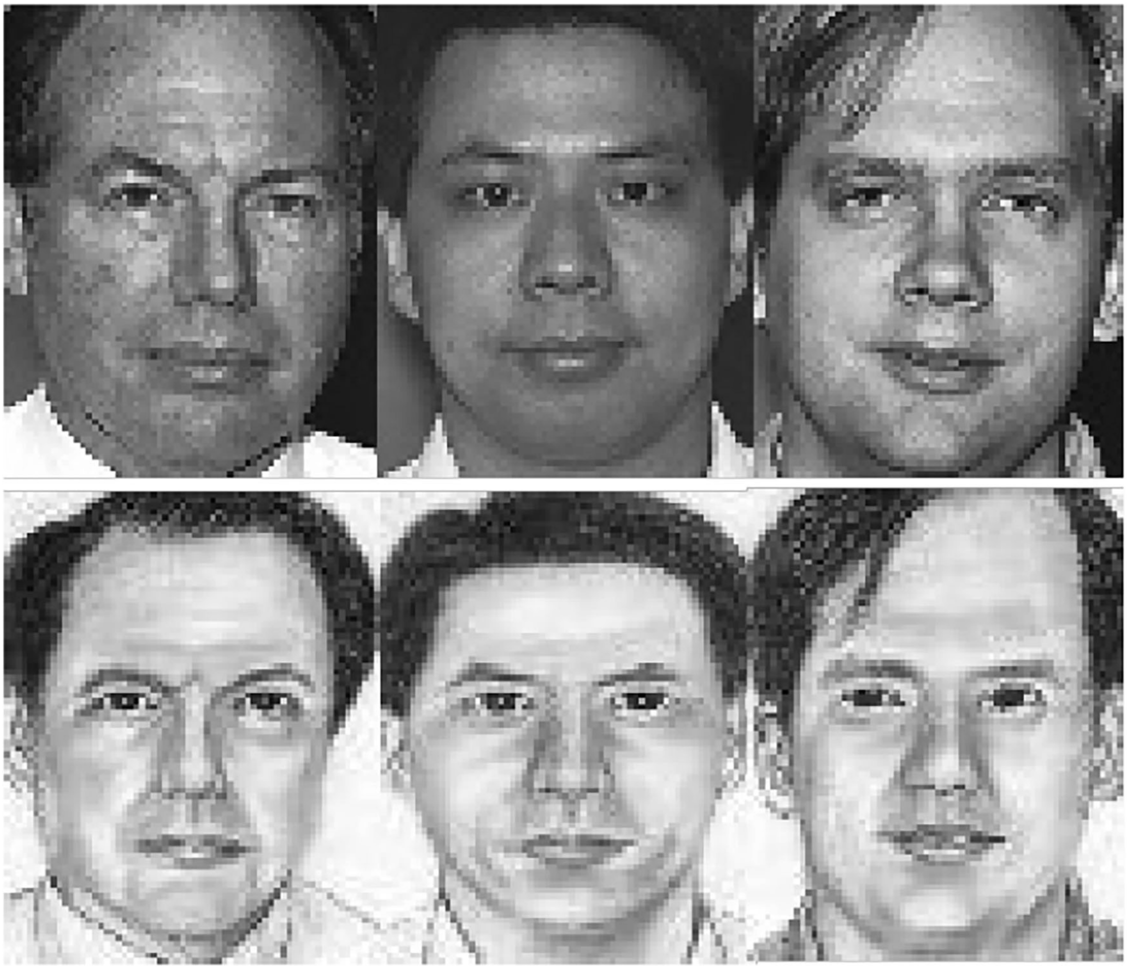


FIGURE 4 | Examples of cropped images from the CUFSF dataset, the top and bottom row are photo and sketch images, respectively.

509 subjects from the MEDS-II dataset[2], and 813 subjects from the LFW dataset.[3] The S3 setup is based on the CUFSF dataset and follows the protocols by Mittal et al. (2015). The training set consists of 500 randomly selected subjects, and the test set contains rest 694 subjects. When tested, the 694 sketch images form the probe set and 694 photo images form the gallery set. All approaches are calculated over 5 train/test set splits. The S4 setup is based on the PRIP-VSGC dataset and follows the protocols by Mittal et al. (2015). The training set consists of 45 randomly selected subjects, and the test set contains the rest 75 subjects. All approaches are calculated over 5 train/test set splits. **Table 1** details four experimental setups.

4.2. Implementation Details

Sketch and photo images are aligned, cropped, and reshaped to 256 × 256 by using the MTCNN (Zhang et al., 2016). **Figures 3**, **4** depict representative cropped images from the UoM-SGFSv2 and CUFSF dataset. Representative data augmentation techniques including random cropping, filling, horizontal flipping, and normalization are employed in the training stage. Specifically, we first pad the images on all sides with the 10 value, next crop the given image at a random location to 256 × 256, then horizontally flip the images randomly with a probability of 0.5, finally normalize the images with mean value of $(0.5, 0.5, 0.5)$ and SD value of $(0.5, 0.5, 0.5)$. Adam optimizer (Kingma and

TABLE 2 | Results of the CTMAN, w/o GeM, w/o CTM, w/o CTM&MLS, and baseline on the S1 setup.

Methods	Rank-1 (%)	Rank-10 (%)	Rank-50 (%)
CTMAN	78.67	96.00	99.20
w/o GeM	74.53	96.00	99.33
w/o CTM	76.67	95.60	99.33
w/o CTM&MLS	57.47	87.47	95.73
baseline	54.93	86.93	95.33

[2] Available at: http://www.nist.gov/itl/iad/ig/sd32.cfm.

[3] Available at: http://vis-www.cs.umass.edu/lfw/.

Ba, 2014) with $(\beta_1, \beta_2) = (0.5, 0.999)$ is utilized to optimize the MAE learning network, the learning rate is set to 0.0001. The total training episode is set to 60, the training episode T is set to 100, the training episode classes b is set to 28, and the memory size M is set to 512. The trade-off parameter λ is set to 0.5 empirically. The first m episode is set to 30.

TABLE 3 | Results of the CTMAN, w/o GeM, w/o CTM, w/o CTM&MLS, and baseline on the S2 setup.

Methods	Rank-1 (%)	Rank-10 (%)	Rank-50 (%)
CTMAN	85.73	98.13	99.33
w/o GeM	82.13	98.13	99.60
w/o CTM	85.33	98.00	98.93
w/o CTM&MLS	70.80	93.07	97.60
baseline	69.20	93.07	98.00

TABLE 4 | Results of the CTMAN, w/o GeM, w/o CTM, w/o CTM&MLS, and baseline on the S3 setup.

Methods	Rank-1 (%)	Rank-10 (%)	Rank-50 (%)
CTMAN	90.06	98.70	99.39
w/o GeM	85.85	98.65	99.34
w/o CTM	89.25	98.73	99.36
w/o CTM&MLS	83.86	97.90	99.34
baseline	80.66	97.35	99.45

4.3. Results and Analysis

4.3.1. Ablation Study

To verify the effectiveness of each component of the proposed CTMAN, we compare CTMAN with w/o GeM, w/o CTM, w/o CTM&MLS, and baseline approach. To verify the effectiveness of the GeM pooling layer, for w/o GeM, the GeM pooling layer is replaced by the traditional maxpooling layer. To verify the effectiveness of the cross task memory mechanisms, for w/o CTM, in each training episode, the cross task modality alignment loss computed by the cross task support feature set is removed, and the loss function is set to Equation (9). To verify the effectiveness of the meta learning training episode strategy, for w/o CTM&MLS, on the basis of w/o CTM, the meta learning training episode strategy and corresponding loss are further removed, it uses the traditional batch training process, and extracts features by MAE learning, then a batch norm layer and linear layer transform the feature into a vector of class logits, the loss is set to cross-entropy loss, the batch size is set to 28,

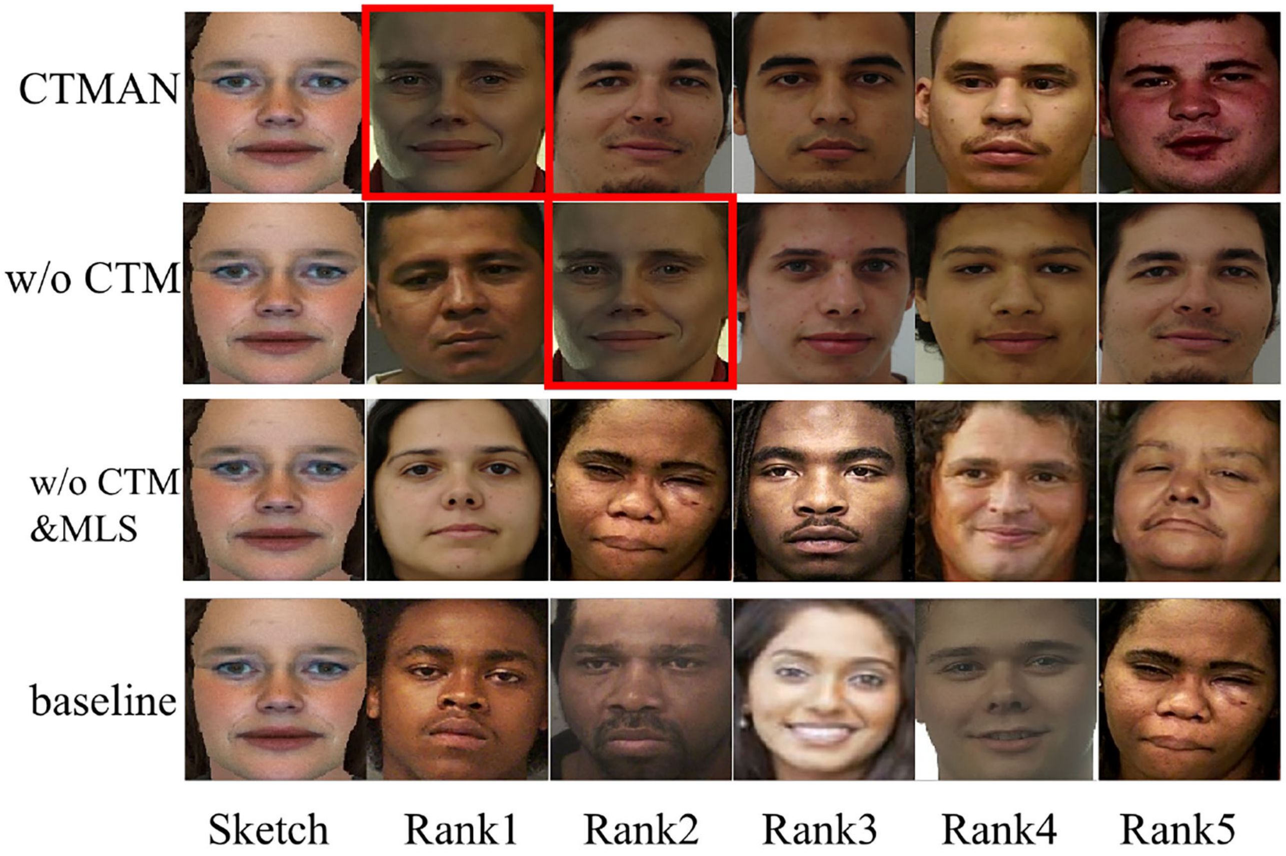


FIGURE 5 | The top five matching photos of CTMAN, w/o CTM, w/o CTM&MLS, and baseline on the S1 setup, images in red box are the groundtruth.

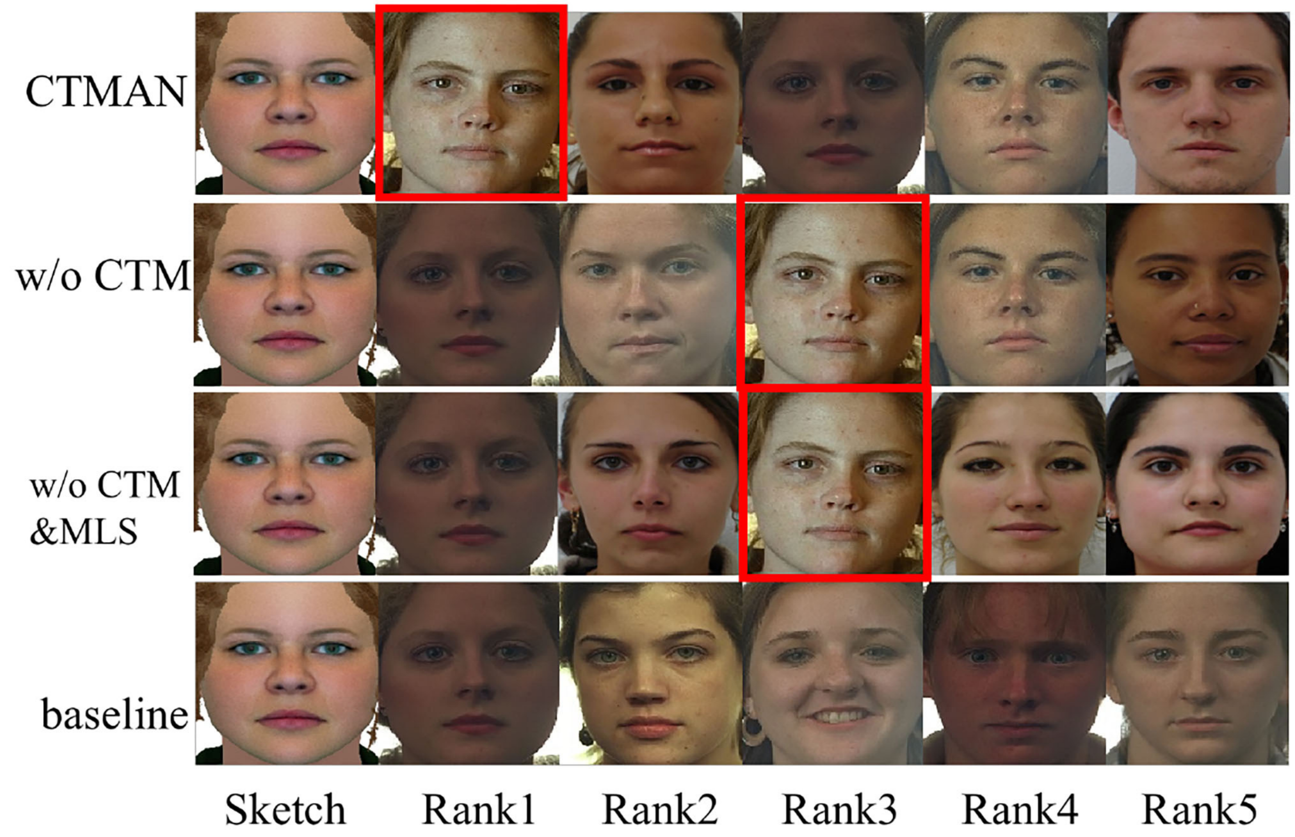


FIGURE 6 | The top five matching photos of CTMAN, w/o CTM, w/o CTM&MLS, and baseline on the S2 setup, images in red box are the groundtruth.

and the epoch is set to 60. For the baseline, on the basis of w/o CTM&MLS, the MAE learning is further removed, it extracts features by the ResNet50 network pretrained on ImageNet. Note that each method uses the same parameter settings and partition protocols to make experiments fair.

Tables 2–4 show the performance of the CTMAN, w/o GeM, w/o CTM, w/o CTM&MLS, and baseline on the S1, S2, and S3 setup. **Figures 5–7** visualize the top five matching photos of CTMAN, w/o CTM, w/o CTM&MLS and baseline on the S1, S2, and S3 setup, respectively, images in red box are the groundtruth. As shown in **Figures 5–7**, we visualize the effect of the four approaches to evaluate our CTMAN's recognition performance intuitively. For each figure, the first line shows the matching results for the proposed method, the second line depicts the results of the w/o CTM, the third line depicts the results of the w/o CTM&MLS, and the final line depicts the result of the baseline. Results show that all methods are lower on the more difficult S1 setup than the S2 setup, and our CTMAN outperforms the w/o GeM, w/o CTM, w/o CTM&MLS, and baseline in three datasets, demonstrating the effectiveness of each contribution of the CTMAN. Compared to baseline, w/o CTM&MLS gains higher performance, illustrating the effectiveness of the MAE learning. Compared to w/o CTM&MLS, w/o CTM gains higher accuracy, illustrating the effectiveness of the meta learning training episode strategy. Compared to w/o CTM, CTMAN gains better performance, demonstrating the effectiveness of the cross task memory mechanism. Compared to w/o GeM, CTMAN gains higher accuracy, illustrating the effectiveness of the GeM pooling layer.

4.3.2. Comparison to the State-of-the-Art Methods

For the first two setup, performance of the CTMAN with the CTMAN*, CTMAN-ResNet18, PCA (Turk, 1991), ET(+PCA) (Tang and Wang, 2004), EP(+PCA) (Galea and Farrugia, 2015), LLE(+PCA) (Chang et al., 2004), CBR (Hu et al., 2013), D-RS (Klare and Jain, 2015), CBR+D-RS (Klare and Jain, 2015), LGMS (Galea and Farrugia, 2016), HAOG (Galoogahi and Sim, 2012), VGG-Face (Parkhi et al., 2015), DEEPS (Galea and Farrugia, 2018), Xu's (Xu et al., 2021), DLFace (Peng et al., 2019), SSR (Peng et al., 2021), and DAEN (Guo et al., 2021) methods are reported in **Tables 5**, **6**. The performance of these compared approaches is directly from Galea and Farrugia (2018), Xu et al. (2021), Peng et al. (2019), Peng et al. (2021), and Guo et al. (2021). The extended gallery set in Galea and Farrugia (2018) consists of part images of the FEI, MEDS-II, Multi-PIE (Gross et al., 2010), and FRGC v2.0[4] datasets, these images are frontal and have high quality. Our extended gallery set (Galea and Farrugia, 2018) consists of part images of the FEI, MEDS-II, and LFW

[4] http://www.nist.gov/itl/iad/ig/frgc.cfm.

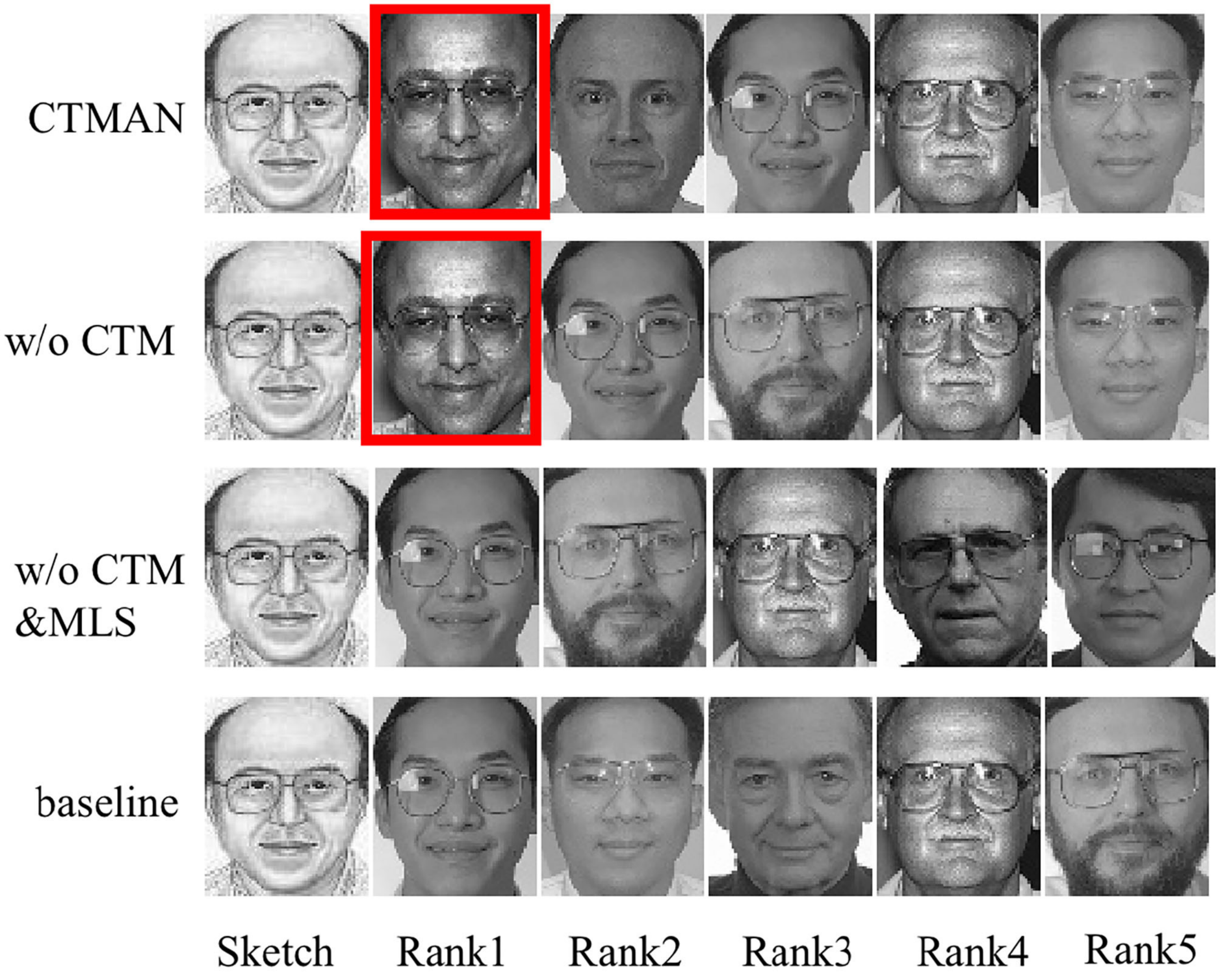


FIGURE 7 | The top five matching photos of CTMAN, w/o CTM, w/o CTM&MLS, and baseline on the S3 setup, images in red box are the groundtruth.

datasets, images of the LFW dataset are captured under the unconstrained environment, they may not be the best replaced images for the Multi-PIE and FRGC datasets. Since images of FRGC and Multi-PIE are not available, Peng et al. (2019) extend the gallery set by 1,180 photos of the XM2VTS dataset (Messer, 1999), 3,098 photos of CAS-PEAL dataset (Gao et al., 2008a), and 3,000 photos of LFW dataset, here we further extend the gallery set in Section 4.1 to 2,277 subjects, the 2,277 subjects include 150 test subjects, 1,521 subjects from the former extend gallery set in Section 4.1 (199 subjects from the FEI dataset, 509 subjects from the MEDS-II dataset, and 813 subjects from the LFW dataset), 188 subjects from the CUHK dataset (Wang and Tang, 2009), 123 subjects from the AR dataset (Martinez and Benavente, 1998), 295 subjects from the XM2VTS dataset (Messer, 1999), selected photos in CUHK, AR, and XM2VTS datasets are taken from the constrained environment. **Figure 8** shows several cropped images in the following datasets: (top row) sketch in UoM-SGFSv2, photo in UoM-SGFSv2, FEI, MEDS-II, LFW, (last row) Multi-PIE, FRGC v2.0, CUHK, AR, and XM2VTS. As shown in **Figure 8**, selected photos in CUHK, AR, and XM2VTS datasets are frontal and have neutral expressions and with minimal shadows and occlusions, these images may be the better replacement for the Multi-PIE and FRGC datasets.

The CTMAN* means CTMAN tested on the extended gallery set with 2,277 photos. For CTMAN-ResNet18, it replaces the ResNet50 backbone of the CTMAN by ResNet18 backbone. The VGG-Face and PCA are traditional face recognition methods, ET(+PCA), EP(+PCA), and LLE(+PCA) are intra-modality methods, the LGMS, HAOG, DEEPS, Xu's, DLFace, SSR, and DAEN are inter-modality methods. As shown in **Tables 5**, **6**, the proposed CTMAN achieves the best performance, it outperforms the second 8% and 12% on rank-1, suggesting the superior performance of CTMAN in the challenging UoM-SGFSv2 dataset. Compared to the UoM-SGFSv2 set B, the

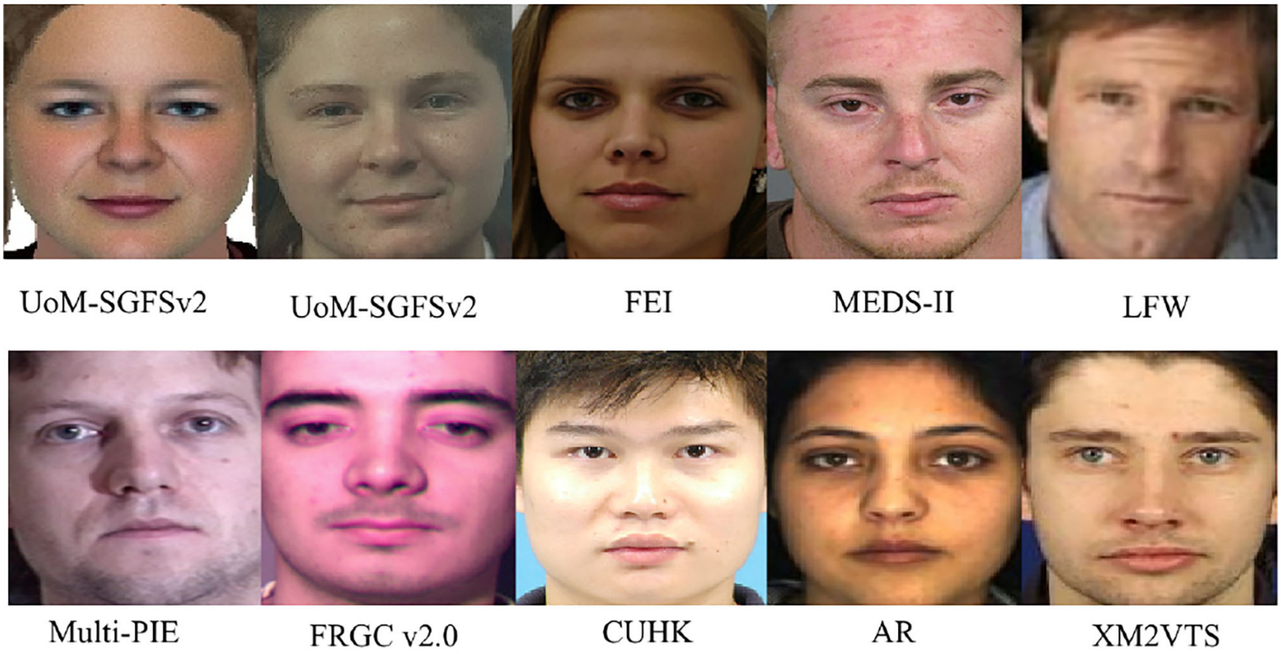


FIGURE 8 | Examples of cropped images in the following datasets: (top row) sketch in UoM-SGFSv2, photo in UoM-SGFSv2, FEI, MEDS-II, LFW , (last row) Multi-PIE, FRGC v2.0, CUHK, AR, and XM2VTS.

TABLE 5 | Comparison experiment results on the S1 setup.

Type	Methods	Rank-1 (%)	Rank-10 (%)	Rank-50 (%)
Face recognition methods	VGG-Face	9.33	31.07	59.73
	PCA	2.80	8.40	17.73
Intra-modality methods	ET+PCA	8.40	30.00	54.53
	EP+PCA	12.53	35.60	62.80
	LLE+PCA	6.93	24.67	43.60
Inter-modality methods	LGMS	21.87	51.20	72.40
	CBR	5.73	18.80	43.33
	D-RS	22.13	49.33	69.87
	D-RS+CBR	25.87	56.00	76.27
	HAOG	13.60	37.33	52.67
	DEEPS	31.60	66.13	86.00
	Xu's	62.00	92.30	-
	DLFace	64.80	92.13	-
	SSR	70.16	94.60	-
	DAEN	68.53	92.40	97.47
Proposed	CTMAN-ResNet18	76.67	96.53	98.93
	CTMAN*	77.60	96.00	99.07
	CTMAN	78.67	96.00	99.20

TABLE 6 | Comparison experiment results on the S2 setup.

Type	Methods	Rank-1 (%)	Rank-10 (%)	Rank-50 (%)
Face recognition methods	VGG-Face	16.13	48.00	72.80
Intra-modality methods	ET+PCA	12.13	39.07	63.47
	EP+PCA	15.20	48.27	70.00
	LLE+PCA	10.53	31.60	53.53
Inter-modality methods	LGMS	21.87	51.2	72.40
	CBR	7.60	25.47	48.27
	D-RS	40.80	70.80	86.40
	D-RS+CBR	42.93	75.87	90.13
	HAOG	21.60	42.27	57.07
	DEEPS	52.17	82.67	94.00
	Xu's	76.00	95.8	-
	DLFace	72.53	94.8	-
	SSR	73.83	95.10	-
	DAEN	74.00	95.20	99.07
Proposed	CTMAN*	85.60	98.13	99.20
	CTMAN	85.73	98.13	99.33

The CTMAN means CTMAN tested on the extended gallery set with 2277 photos.*

accuracy of all approaches are lower on the challenging UoM-SGFSv2 set A. Performance of the inter-modality methods is generally better than the intra-modality methods on the UoM-SGFSv2 set A and B because the performance of intra-modality is a traditional simple method and depends on the quality of the generated image heavily, resulting in degradation of the performance. Despite the VGG-Face method achieving state-of-the-art performance for traditional face recognition, it generally yields poor performance for sketch face recognition in the lower

TABLE 7 | Comparison experiment results on the S3 setup.

Type	Methods	Rank-1 (%)
Intra-modality methods	MWF	74.00
	Fast-RSLCR	75.94
	Wan's	70.00
Inter-modality methods	Transfer deep feature learning	72.38
	CMML	75.94
	CDFL	81.30
	CMTDML	83.86
Proposed	CTMAN	90.06

TABLE 8 | Comparison experiment results on S4 setup.

Type	Methods	Rank-10%
traditional methods	SSD	45.30
	Attribute	53.10
deep learning methods	Transfer Learning	52.00
	DAEN	63.20
proposed	CTMAN	65.33

ranks, demonstrating the challenging modality gap between photos and sketches. In each batch, training sketch and photo images are randomly selected from the training set, they may not be paired. Instead, we randomly select sketch and photo images paired in each episode. Furthermore, the batch size and epoch used in the two methods were different, these differences may cause the performance gap. Compared to CTMAN, CTMAN* shows comparable performance and outperforms other compared methods, demonstrating the robustness of the CTMAN. CTMAN-ResNet18 outperforms DAEN by a large margin, demonstrating the effectiveness of the proposed method.

For the third setup, the performance of the CTMAN with the MWF (Zhou et al., 2012), Fast-RSLCR (Wang N. et al., 2018), Wan's (Wan and Lee, 2019), CMML (Mignon and Jurie, 2012), CDFL (Jin et al., 2015), Transfer Deep Feature Learning (Wan et al., 2019), and CMTDML (Feng et al., 2019) methods are reported in **Table 7**. Performance of these compared approaches are directly from Feng et al. (2019). Fast RSLCR, MWF, Wan's are intra-modality methods while CDFL, CMML, Transfer Deep Feature Learning, and CMTDML are representative inter-modality method. As shown in **Table 7**, the proposed CTMAN achieves the highest performance, it outperforms the second by nearly 6% on rank-1, which shows the robustness of CTMAN on the CUFSF dataset.

For the fourth setup, the performance of the CTMAN with the SSD (Mittal et al., 2014), Attribute (Mittal et al., 2017), Transfer Learning (Mittal et al., 2015), and DAEN (Guo et al., 2021) methods are reported in **Table 8**. The performance of these compared approaches are directly from Mittal et al. (2015), Mittal et al. (2017), and Guo et al. (2021). The SSD and Attribute are traditional methods, whereas Transfer Learning and DAEN are deep learning methods. As shown in **Table 8**, the proposed CTMAN achieves the highest performance, it outperforms the second by nearly 2% on rank-1, which shows the effectiveness of CTMAN on the PRIP-VSGC dataset.

5. CONCLUSION

In this paper, the CTMAN is proposed for sketch face recognition. By introducing a meta learning training episode strategy, a MAE learning and proposing a cross task memory mechanism, a query feature set, two support feature set and two cross task support feature set and have been extracted to incorporate modal information as well as mimic few-shot tasks, then a cross task modality alignment loss and a modality alignment loss have computed on the above feature set to guide the network to learn discriminative features. Extensive experiments have been conducted on the UoM-SGFSv2, CUFSF, and PRIP-VSGC datasets. Ablation studies have illustrated the effectiveness of the meta training episode strategy, MAE learning, cross task memory mechanism, and cross task modality alignment loss. Comparisons with extensive inter-model and intra-model sketch face recognition approaches have validated the superiority of the CTMAN.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

YG: ideas, formulation, and evolution of overarching research goals and aims, creation and presentation of the published work, and specifically writing the initial draft. LC: provision of study materials, reagents, materials, specifically critical review, commentary, and revision. KD: specifically visualization and data presentation, and specifically critical review. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (62001033 and U20A20163), the Qin Xin Talents Cultivation Program of Beijing Information Science and Technology University (QXTCP A201902 and QXTCPC 202108), and by the General Foundation of Beijing Municipal Commission of Education (KZ202111232049, KM202011232021, and KM202111232014).

REFERENCES

Bhatt, H. S., Bharadwaj, S., Singh, R., and Vatsa, M. (2010). "On matching sketches with digital face images," in *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems*. doi: 10.1109/BTAS.2010.5634507

Chang, H., Yeung, D. Y., and Xiong, Y. (2004). "Super-resolution through neighbor embedding," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2004.1315043

Dhillon, G. S., Chaudhari, P., Ravichandran, A., and Soatto, S. (2019). A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*.

Fan, L., Sun, X., and Rosin, P. L. (2020). "Siamese graph convolution network for face sketch recognition: an application using graph structure for face photo-sketch recognition," in *International Conference on Pattern Recognition*.

Feng, Y., Wu, F., and Huang, Q. (2019). "Cross-modality multi-task deep metric learning for sketch face recognition," in *2019 Chinese Automation Congress*, 2277–2281. doi: 10.1109/CAC48633.2019.8996397

Galea, C., and Farrugia, R. A. (2018). Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning. *IEEE Trans. Inform. Forensics Sec*. 13, 1421–1431. doi: 10.1109/TIFS.2017.2788002

Galea, C., and Farrugia, R. A. (2015). "Fusion of intra- and inter-modality algorithms for face-sketch recognition," in *Computer Analysis of Images and Patterns*, 700–711. doi: 10.1007/978-3-319-23117-4_60

Galea, C., and Farrugia, R. A. (2016). "Face photo-sketch recognition using local and global texture descriptors," in *European Signal Processing Conference*. doi: 10.1109/EUSIPCO.2016.7760647

Galoogahi, H. K., and Sim, T. (2012). "Inter-modality face sketch recognition," in *IEEE International Conference on Multimedia and Expo*. doi: 10.1109/ICME.2012.128

Gao, W., Cao, B., Shan, S., Chen, X., and Zhou, D. (2008a). The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Trans. Syst. Man Cybernet. A* 38, 2277–2281. doi: 10.1109/TSMCA.2007.909557

Gao, X., Zhong, J., Jie, L., and Tian, C. (2008b). Face sketch synthesis algorithm based on e-HMM and selective ensemble. *IEEE Trans. Circ. Syst. Video Technol*. 18, 487–496. doi: 10.1109/TCSVT.2008.918770

Garcia, V., and Bruna, J. (2017). "Few-shot learning with graph neural networks," in *International Conference on Learning Representations*.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image Vis. Comput*. 28, 807–813. doi: 10.1016/j.imavis.2009.08.002

Guo, Y., Cao, L., Chen, C., Du, K., and Fu, C. (2021). Domain alignment embedding network for sketch face recognition. *IEEE Access* 9, 872–882. doi: 10.1109/ACCESS.2020.3047108

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. doi: 10.1109/CVPR.2016.90

Hu, H., Klare, B. F., Bonnen, K., and Jain, A. K. (2013). Matching composite sketches to face photos: a component-based approach. *IEEE Trans. Inform. Forensics Sec*. 8, 191–204. doi: 10.1109/TIFS.2012.2228856

Jiang, L., Zhong, C., Kailun, W., Gang, Z., and Changshui, Z. (2018). "Boosting few-shot image recognition via domain alignment prototypical networks," in *International Conference on Tools with Artificial Intelligence*.

Jin, Y., Lu, J., and Ruan, Q. (2015). Coupled discriminative feature learning for heterogeneous face recognition. *IEEE Trans. Inform. Forensics Sec*. 10, 640–652. doi: 10.1109/TIFS.2015.2390414

Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klare, B., and Jain, A. K. (2015). "Heterogeneous face recognition: matching NIR to visible light images," in *IEEE Conference on International Conference on Pattern Recognition*.

Lin, W.-H., Wu, B.-H., and Huang, Q.-H. (2018). "A face-recognition approach based on secret sharing for user authentication in public-transportation security," in *IEEE International Conference on Applied System Innovation*. doi: 10.1109/ICASI.2018.8394545

Martinez, A., and Benavente, R. (1998). *The AR Face Database*. CVC technical report.

Méndez-Vázquez, H., Becerra-Riera, F., Morales-Gonzalez, A., Lopez-Avila, L., and Tistarelli, M. (2019). "Local deep features for composite face sketch recognition," in *International Workshop on Biometrics and Forensics*, 1–6. doi: 10.1109/IWBF.2019.8739212

Messer, K. (1999). "XM2VTSDB: the extended M2VTS database," in *Audio and Video Based Biometric Person Authentication*, 72–77.

Mignon, A., and Jurie, F. (2012). "CMML: a new metric learning approach for cross modal matching," in *Asian Conference on Computer Vision*.

Mittal, P., Jain, A., Goswami, G., Singh, R., and M. Vatsa. (2014). "Recognizing composite sketches with digital face images via ssd dictionary," in *IEEE International Joint Conference on Biometrics*, 1–6.

Mittal, P., Jain, A., Goswami, G., Vatsa, M., and Singh, R. (2017). Composite sketch recognition using saliency and attribute feedback. *Inform. Fusion* 33, 86–99. doi: 10.1016/j.inffus.2016.04.003

Mittal, P., Vatsa, M., and Singh, R. (2015). "Composite sketch recognition via deep network - a transfer learning approach," in *International Conference on Biometrics*, 251–256. doi: 10.1109/ICB.2015.7139092

Parkhi, O., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognitions," in *British Machine Vision Conference*. doi: 10.5244/C.29.41

Peng, C., Wang, N., Li, J., and Gao, X. (2019). DLFACE: deep local descriptor for cross-modality face recognition. *Pattern Recogn*. 90, 161–171. doi: 10.1016/j.patcog.2019.01.041

Peng, C., Wang, N., Li, J., and Gao, X. (2021). Soft semantic representation for cross-domain face recognition. *IEEE Trans. Inform. Forensics Secur*. 16, 346–360. doi: 10.1109/TIFS.2020.3013209

Radenovic, F., Tolias, G., and Chum, O. (2017). Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 1655–1668. doi: 10.1109/TPAMI.2018.2846566

Rallings, C., Thrasher, M., Gunter, C., Phillips, P. J., and Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image Vision Comput. J*. 16, 295–306. doi: 10.1016/S0262-8856(97)00070-X

Robinson, J., Chuang, C., Sra, S., and Jegelka, S. (2021). "Contrastive learning with hard negative samples," in *International Conference on Learning Representations*.

Snell, J., Swersky, K., and Zemel, R. (2017). "Prototypical networks for few-shot learning," in *Conference and Workshop on Neural Information Processing Systems*.

Tang, X., and Wang, X. (2004). Face sketch recognition. *IEEE Trans. Circ. Syst. Video Technol*. 14, 50–57. doi: 10.1109/TCSVT.2003.818353

Turk, M. (1991). Eigenfaces for recognition. *J. Cogn. Neurosci*. 3, 71–86. doi: 10.1162/jocn.1991.3.1.71

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 3630–3638.

Wan, W., Gao, Y., and Lee, H. (2019). Transfer deep feature learning for face sketch recognition. *Neural Comput. Appl*. 31, 9175–9184. doi: 10.1007/s00521-019-04242-5

Wan, W., and Lee, H. J. (2019). "Generative adversarial multi-task learning for face sketch synthesis and recognition," in *2019 IEEE International Conference on Image Processing*, 4065–4069. doi: 10.1109/ICIP.2019.8803617

Wang, J., Zhu, Z., Li, J., and Li, J. (2018). "Attention based siamese networks for few-shot learning," in *IEEE 9th International Conference on Software Engineering and Service Science*, 551–554. doi: 10.1109/ICSESS.2018.8663732

Wang, N., Gao, X., and Li, J. (2018). Random sampling for fast face sketch synthesis. *Pattern Recogn*. 76, 215–227. doi: 10.1016/j.patcog.2017.11.008

Wang, N., Gao, X., Sun, L., and Li, J. (2017a). Bayesian face sketch synthesis. *IEEE Trans. Image Process*. 26, 1264–1274. doi: 10.1109/TIP.2017.2651375

Wang, N., Zha, W., Li, J., and Gao, X. (2017b). Back projection: an effective postprocessing method for GAN-based face sketch synthesis. *Pattern Recogn. Lett*. 107, 59–65. doi: 10.1016/j.patrec.2017.06.012

Wang, X., Girshick, R., Gupta, A., and He, K. (2017c). "Non-local neural networks," in 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT). doi: 10.1109/CVPR.2018.00813

Wang, X., and Tang, X. (2009). Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell*. 31, 1955–1967. doi: 10.1109/TPAMI.2008.222

Wang, X., Zhang, H., Huang, W., and Scott, M. R. (2019). "Cross-batch memory for embedding learning," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE). doi: 10.1109/CVPR42600.2020.00642

Xu, J., Xue, X., Wu, Y., and Mao, X. (2021). Matching a composite sketch to a photographed face using fused hog and deep feature models. *Visual Comput*. 37, 1–12. doi: 10.1007/s00371-020-01976-5

Ye, M., Lan, X., Wang, Z., and Yuen, P. C. (2020). Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Trans. Inform. Forensics Sec*. 15, 407–419. doi: 10.1109/TIFS.2019.2921454

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. H. (2021). Deep learning for person re-identification: a survey and outlook. *arXiv preprint arXiv:2001.04193*.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett*. 23, 1499–1503. doi: 10.1109/LSP.2016.2603342

Zhang, L., Lin, L., Wu, X., Ding, S., and Zhang, L. (2015). "End-to-end photo-sketch generation via fully convolutional representation learning," in *5th ACM on International Conference on Multimedia Retrieval*, 627–634. doi: 10.1145/2671188.2749321

Zhong, Z., Zheng, L., Luo, Z., Li, S., and Yang, Y. (2019). "Invariance matters: exemplar memory for domain adaptive person re-identification," *In IEEE Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2019.00069

Zhou, H., Kuang, Z., and Wong, K. K. (2012). "Markov weight fields for face sketch synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1091–1097.

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership