frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 06 April 2023
DOI 10.3389/fphy.2023.1180100

OPEN ACCESS

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

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

*CORRESPONDENCE
Yafei Zhang,
zyfeimail@kust.edu.cn

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

RECEIVED 05 March 2023
ACCEPTED 23 March 2023
PUBLISHED 06 April 2023

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

Infrared and visible image fusion with edge detail implantation

Junyu Liu, Yafei Zhang* and Fan Li

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

Infrared and visible image fusion aims to integrate complementary information from the same scene images captured by different types of sensors into one image to obtain a fusion image with richer information. Recently, deep learning-based infrared and visible image fusion methods have been widely used. However, it is still a difficult problem how to maintain the edge detail information in the source images more effectively. To address this problem, we propose a novel infrared and visible image fusion method with edge detail implantation. The proposed method no longer improves the performance of edge details in the fused image through making the extracted features contain edge detail information like traditional methods, but by processing source image information and edge detail information separately, and supplementing edge details to the main framework. Technically, we propose a two-branch feature representation framework. One branch is used to directly extract features from the input source image, while the other is utilized to extract features of edge map. The edge detail branch mainly provides edge detail features for the source image input branch, ensuring that the output features contain rich edge detail information. In the fusion of multi-source features, we respectively fuse the source image features and the edge detail features, and use the fusion results of edge details to guide and enhance the fusion results of source image features so that they contain richer edge detail information. A large number of experimental results demonstrate the effectiveness of the proposed method.

KEYWORDS

infrared and visible image fusion, edge detail implantation, information compensation, dual branch network, end-to-end network

1 Introduction

Due to the different imaging mechanisms, two types of images for the same scene often carry a large amount of complementary information. If these complementary information can be integrated into one image, it will help improve the comprehensiveness and accuracy of the image to describe the scene, which is conducive to the development of subsequent tasks. To this end, infrared and visible image fusion technology has been proposed and widely applied to computer vision fields with different tasks, such as object detection [1], face recognition [2], video surveillance [3] and so on.

In recent years, with the rapid development of deep learning, the research of fusion methods among diverse modal information has made significant progress [4-8]. As an important branch in the field of image fusion, infrared and visible image fusion has attracted the attention of researchers, and a series of effective methods have been proposed. These methods can be roughly divided into methods based on multi-scale transformations [9-11], sparse modeling [12-15], and deep learning [16, 17]. Multi-scale transformation based methods include pyramid transform [9], DWT [18], Contourlet

transform (CT) [19], non-subsampled contourlet transform (NSCT) methods[20], etc. This kind of methods cannot achieve sparse expression of image because they use artificially constructed basis functions to represent images, limiting the visual quality improvement of fused images. Methods based on sparse modeling can solve the above problems by using an over-complete dictionary to represent images. However, these methods are difficult to mine the statistical information from large-scale training samples in an effective way, which limits the further improvement of their expression ability.

Among deep learning-based fusion methods, CNN-based methods are most common. At present, there are CNN-based infrared and visible image fusion as Cross-UNet-based [21], ResNet-based [22], GAN-based [23], Encoder-Decoder-based methods [24], etc. In view of the fact that CNNs cannot capture features over long distances, transformer-based infrared and visible image fusion method was proposed. However, since Transformer is designed based on attention mechanism, it has certain limitations in mining detailed information at the edges of the image. To solve the above problems, this paper proposes an infrared and visible image fusion method with edge detail implantation. In terms of feature extraction, the proposed method consists of two feature extraction branches: one is the feature extraction branch based on Transformer, and the other is the edge detail feature extraction branch based on CNN. The former takes infrared and visible source images as input, and the latter takes edge details detected from the source images as input. Information extracted by the latter is fed back to the former to compensate for the limitations of the transformer in extracting features.

In the feature implantation from CNN branch to the Transformer branch, an effective feature implantation method based on attention mechanism is designed, which not only considers the role of common information between different features in two branches, but also the complementary features extracted by CNN branch, realizing effective transmission of CNN features to Transformer branch. In terms of feature fusion, the features extracted from CNN branch and Transformer branch are fused respectively, and the fusion features of the CNN branch are used to guide the fusion features of the Transformer branch, so as to realize the fusion feature transfer from the CNN branch to the Transformer branch. It further compensates the shortcomings of the Transformer in feature extraction. The above method not only combine the advantages of CNN and Transformer in feature extraction into the whole framework, but also effectively enhances the representation ability of edge details, thereby improving the visual quality of the fusion results. In summary, the main contributions of this paper are as follows.

1) A method of infrared and visible image fusion with edge detail information implantation is proposed. This method uses two different branches based on Transformer and CNN to extract features from the input source images and the edge maps of the source images, and implants features extracted by CNN into the Transformer branch to make up for the shortcomings of Transformer in extracting edge details.
2) Based on attention mechanism, an information implantation method is designed, which realizes the injection of CNN branch information into Transformer, effectively making up for the shortcomings of Transformer in extracting features. In addition, the proposed method fuses the features obtained by CNN and Transformer branches respectively, and uses the fusion results of CNN branch to guide the fusion results of Transformer, further maintaining the edge details in fusion results.
3) In order to ensure that the features used to reconstruct fusion result are rich in edge details, we introduce an edge reconstruction block, and use edge detail information of the target image (fusion result) as a constraint to make the reconstruction result consistent with the target image, so as to ensure that feature to reconstruct the fusion result contains relevant information about edge details.

2 Related works

2.1 Infrared-visible image fusion

Infrared and visible image fusion is an important branch of fusion field. According to the previous introduction, current infrared and visible image fusion methods can be divided into fusion methods based on multi-scale transformation [25], sparse representation [26-29], and deep learning [30-32]. Multi-scale transformation based methods usually perform multi-scale decomposition of the input source image first, then fuse the decomposed coefficients, and finally apply the corresponding multi-scale inverse transformation to the fusion result to reconstruct the fusion image. These methods are simple to implement and has good stability. However, due to the use of fixed bases to represent the image information, their sparse expression ability is weak, which limits the improvement of fusion performance. Methods based on sparse representation were more popular 10 years ago. These methods can represent source images in a sparser way, obtaining better fusion performance than the former. However, these methods are difficult to mine statistical characteristics of features from large-scale training samples in an effective way, and thus still have limitations in representing image information.

In recent years, deep learning has been widely used in various image fusion tasks due to its powerful feature extraction and representation capabilities [33-38] without manually designing features and fusion strategies. In particular, Li et al. [31] proposed the DenseFuse infrared-visible image fusion framework, which combined the shallow and deep features of network by using dense blocks in the encoding process to extract richer source image features. In order to improve the fusion performance, Ma et al. [39] proposed a dual-confrontation DDcGAN fusion framework to further improve the performance of FusionGAN [23] when fusing infrared and visible images. Additionally, to more effectively maintain the edge details of the source image, Zhao et al. [40] used different encoders to extract high-frequency detail information and low-frequency information from the source images separately. Li et al. [22] proposed to use the detail preservation loss function and feature enhancement loss function in the residual structure network, combining with the two-stage training strategy to ensure that the fusion results contain rich detail information

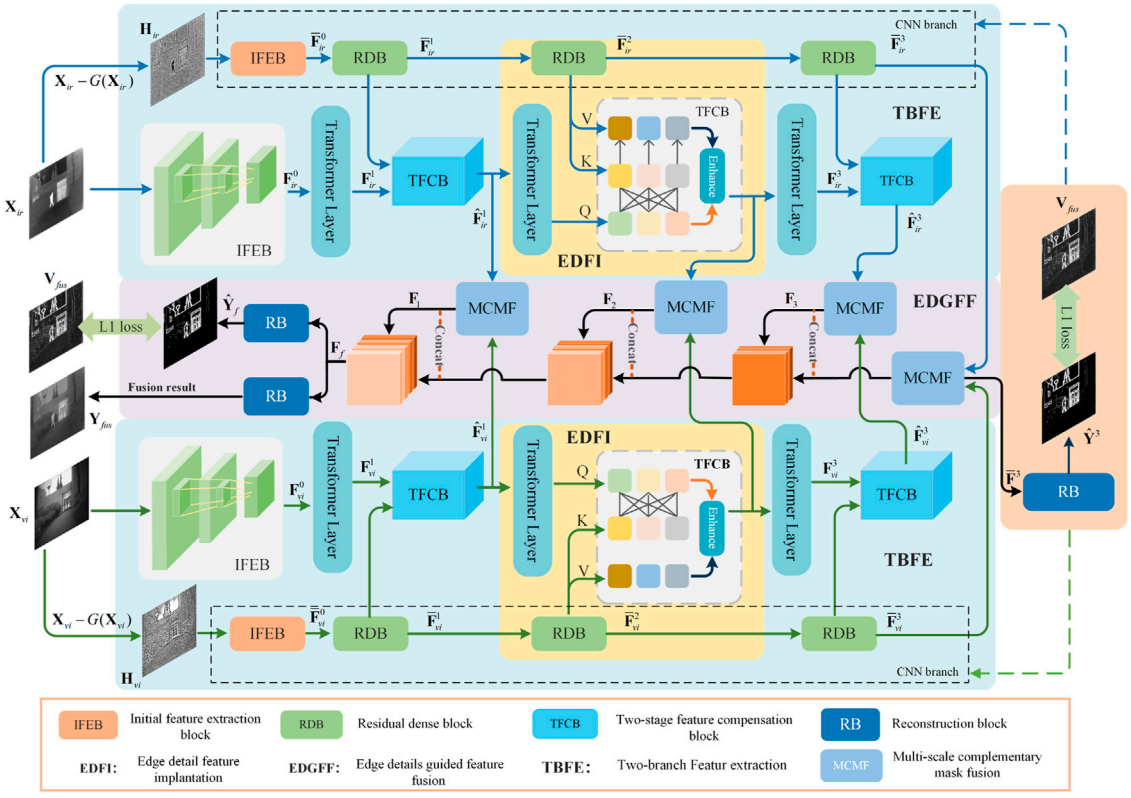

FIGURE 1
Overall framework of the proposed method.

and significant information. Although the above CNN-based methods have achieved certain performance, they are still insufficient in maintaining edge detail information. In addition, CNNs cannot mine the relationship between features over long distances, which limits the further improvement of their performance.

2.2 Transformer based image fusion

Thanks to its excellent long-distance modeling capabilities, Transformers [41] has attracted the attention of researchers in image fusion. In particular, in order to establish the global dependence of image features, Ma et al. [42] proposed a residual fusion framework based on SwinTransformer for infrared and visible image fusion. This framework abandons traditional convolution operations and adopts an attention-based network structure. At the same time, a fusion strategy based on L1 norm is designed, which further improves the fusion quality. In order to obtain high-quality pan-sharpened remote sensing images, Bandara et al. [43] proposed a new hyperTransformer framework. This method transfers the high-resolution texture information in PAN images to LR-HSI image features by attention mechanism, avoiding the image spatial and spectral distortion caused by traditional fusion methods.

In order to combine the respective advantages of CNN and Transformer, Vs. et al. [44] proposed a transformer-based encoding and decoding structure, and used the dual-branch structure of CNN and Transformer to fuse image features. Although this method can obtain satisfactory fusion results, it does not consider the problem of maintaining the edge details of source images, resulting in the loss of source image detail information. Li et al. [45] combined the local features of the convolutional network with the global features of the transformer by alternately using CNN and Transformer in the network, overcoming the shortcomings of a single network and improving the visual quality of the fused image. Based on the multi-scale feature pyramid theory, Park et al. [46] proposed an image fusion method for dual-modality transformers. This method mines the complementary information between source images by estimating the non-correlated mapping relationship between features of the source images, so as to improve the extracted feature quality of the source images. Although the above methods have achieved a certain degree of performance improvement, they does not fully consider the problem of maintaining the edge detail of the source images, which still remains large improvement space of visual effect. Different from the above methods, this paper uses two parallel feature extraction branches, Transformer and CNN, to extract the features of input source images and edge details, respectively, and implant the features extracted by CNN into the Transformer branch. This method can not only effectively integrate the respective advantages of CNN and Transformer, but also avoid the loss of edge detail information.

3 Proposed method

3.1 Overview

The overall framework of the proposed method is shown in Figure 1. It consists of three parts: two-branch feature extraction (TBFE), edge detail feature implantation (EDFI), and edge detail guided feature fusion (EDGFF). TBFE is mainly used to obtain the features and edge details of the source images. In this process, we use Transformer-based network to extract source image features, while utilize CNN network to the extract edge detail features. EDFI is mainly used to inject edge detail features extracted from the CNN branch into the Transformer branch to make up for the Transformer's shortcomings in extracting edge details. EDGFF uses the fused features of CNN branch to guide the fusion of Transformer branch, further highlighting the edge details in the fusion results.

3.2 Two-branch feature extraction

3.2.1 Transformer feature extraction branch

As shown in Figure 1, we use CNN and Transformer for TBFE, respectively, and the TBFEs for infrared and visible images have the same network structure. Since Transformer network has better feature relationship modeling ability for long-distance and can better describe the relationship between different features, we use Transformer branch to extract source image features. This branch takes the source image $\mathbf{X}_j$ ($j = ir, vi$) as the input, and first uses initial feature extraction block (IFEB) to obtain a shallow multi-channel feature map, which is convenient for subsequent Transformer feature extraction. The extracted features can be represented as:

$$\mathbf{F}_j^0 = f_{ifeb}\left(\mathbf{X}_j\right) \tag{1}$$

where f_{ifeb} denotes the feature extraction operation of IFEB. In this paper, IFEB consists of three 3×3 convolutional layers and a ReLU activation function. We utilize Transformer to extract the global features of the obtained feature $\mathbf{F}_j^0$. For the first transformer layer, its input feature is $\mathbf{F}_j^0$, and the output is expressed as:

$$\mathbf{F}_j^1 = f_{t1}\left(\mathbf{F}_j^0\right) \tag{2}$$

where f_{t1} represents the first Transformer layer feature extraction operation, mainly composed of layer normalization (LN), multi-head self-attention layer (MSA) and multi-layer perceptron (MLP). Correspondingly, for the ith ($i \geq 2$) Transformer layer, its output is $\mathbf{F}_j^i$.

3.2.2 CNN feature extraction branch

Compared with Transformer, CNNs are better at describing underlying visual features such as image structure and texture. Therefore, this paper uses CNN branch to extract features of edge details. In order to obtain the edge detail information from the source images, we perform Gaussian smoothing filtering on the source image $\mathbf{X}_j$ to obtain smooth image, and use the source image information to differ from the smooth image to obtain the edge detail information:

$$\mathbf{H}_j = \mathbf{X}_j - G\left(\mathbf{X}_j\right) \tag{3}$$

where G is the Gaussian blur operation. Edge details obtained in this way contain high-frequency information of the source image, which can effectively depict the edge details. Compared with the gradient map extracted by gradient operator, $\mathbf{H}_j$ contains richer edge detail and texture information. Similar to the Transformer branch, we use IFEB to extract the underlying features of the edge detail map in CNN branch:

$$\bar{\mathbf{F}}_j^0 = f_{ifeb}\left(\mathbf{H}_j\right) \tag{4}$$

Besides, detailed features are further extracted by residual dense block (RDB). For the first RDB, its input is features of the edge detail map $\bar{\mathbf{F}}_j^0$ and the output is $\bar{\mathbf{F}}_j^1$:

$$\bar{\mathbf{F}}_j^1 = f_{rdb1}\left(\bar{\mathbf{F}}_j^0\right) \tag{5}$$

where f_{rdb1} represents the feature extraction operation of the first RDB. In this work, RDB is a feature extraction block composed of three convolutional layers, ReLU activation function and densely connected between them. Correspondingly, for the ith ($i \geq 2$) RDB, its output is $\bar{\mathbf{F}}_j^i$.

3.3 Edge detail feature implantation

In order to make the extracted features in Transformer branch rich in edge detail, a two-stage feature compensation block (TFCB) is proposed, as shown in Figure 2. This module solves the problem that the Transformer branch is difficult to extract edge detail features by implanting local texture details of the image extracted by RDB into the Transformer branch. As for network structure, the module consists of two stages of feature compensation. The feature compensation in the first stage realizes the transmission of information by finding the correlation between $\bar{\mathbf{F}}_j^i$ and $\mathbf{F}_j^i$, and dynamically aggregating features of $\bar{\mathbf{F}}_j^i$ according to the changes of input features. Specifically, $\bar{\mathbf{F}}_j^i$ and $\mathbf{F}_j^i$ are first transformed into three feature spaces $\mathbf{Q}_j^i$, $\bar{\mathbf{K}}_j^i$, and $\bar{\mathbf{V}}_j^i$ by 1×1 convolution.

$$\begin{cases} \mathbf{Q}_j^i = \mathrm{Conv}_{1\times1}\left(\mathbf{F}_j^i\right) \\ \bar{\mathbf{K}}_j^i = \mathrm{Conv}_{1\times1}\left(\bar{\mathbf{F}}_j^i\right) \\ \bar{\mathbf{V}}_j^i = \mathrm{Conv}_{1\times1}\left(\bar{\mathbf{F}}_j^i\right) \end{cases} \tag{6}$$

where $\mathrm{Conv}_{1\times1}$ denotes 1×1 convolution. The first stage feature compensation process can be formulated as:

$$\tilde{\mathbf{F}}_j^i = \mathbf{F}_j^i + softmax\left(\frac{\mathbf{Q}_j^i\left(\bar{\mathbf{K}}_j^i\right)^T}{\sqrt{C}}\right)\bar{\mathbf{V}}_j^i \tag{7}$$

Where C is the dimension of $\bar{\mathbf{K}}_j^i$. The above method achieves information transfer from $\bar{\mathbf{F}}_j^i$ to $\mathbf{F}_j^i$ by using $\bar{\mathbf{F}}_j^i$ to represent $\mathbf{F}_j^i$. However, due to differences between $\bar{\mathbf{F}}_j^i$ and $\mathbf{F}_j^i$, the features re-aggregated based on similarity may lose some details. In order to avoid this problem, this paper introduces the second stage of feature compensation. Specifically, we input the re-aggregated features

$$\mathbf{T}_j^i = softmax\left(\frac{\mathbf{Q}_j^i\left(\bar{\mathbf{K}}_j^i\right)^T}{\sqrt{C}}\right)\bar{\mathbf{V}}_j^i \tag{8}$$

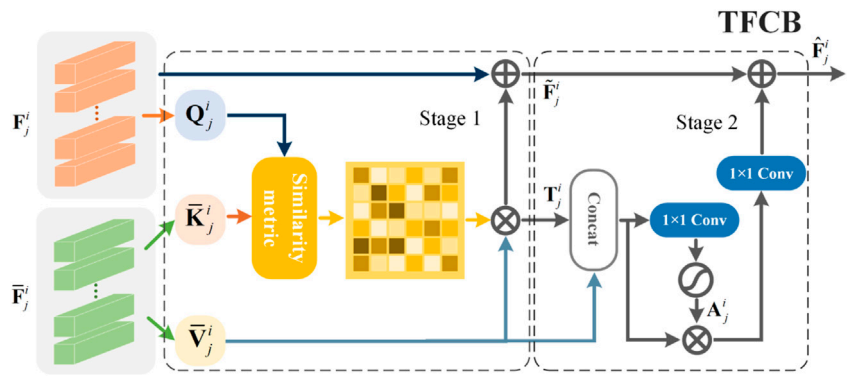

FIGURE 2
Detailed structure of TFCB.

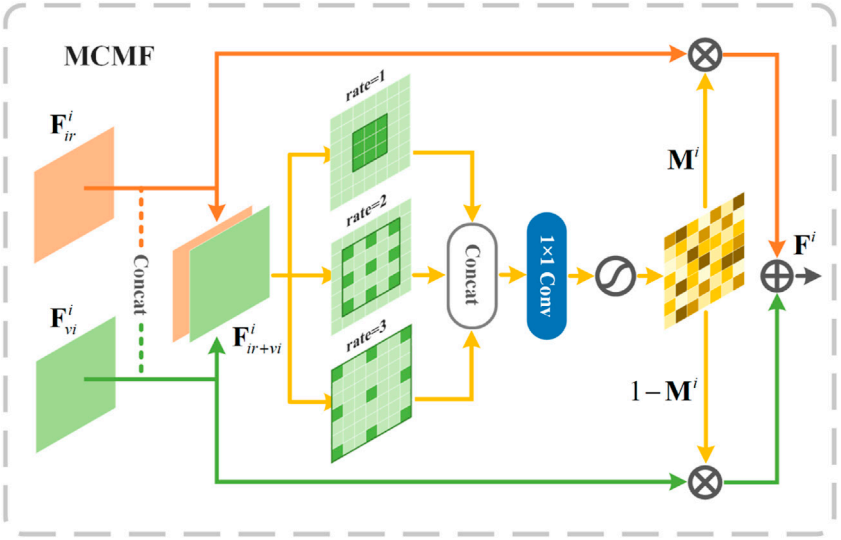

FIGURE 3
Detailed structure of MCMF.

and $\bar{\mathbf{V}}_j^i$ to a small CNN network, and select the activated features of the network through attention map, performing the second information compensation, as shown in Figure 2. To obtain the spatial attention map, we first concatenate $\mathbf{T}_j^i$ and $\bar{\mathbf{V}}_j^i$, and apply 1 × 1 convolution and Sigmoid:

$$\mathbf{A}_j^i = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{concat}\left(\mathbf{T}_j^i, \bar{\mathbf{V}}_j^i\right)\right)\right) \tag{9}$$

where σ represents the Sigmoid activation function. Output features of the ith information compensation module can be represented as:

$$\hat{\mathbf{F}}_j^i = \tilde{\mathbf{F}}_j^i + \mathrm{Conv}_{1\times1}\left(\mathbf{A}_j^i \odot \mathrm{concat}\left(\mathbf{T}_j^i, \bar{\mathbf{V}}_j^i\right)\right) \tag{10}$$

The two-stage feature compensation strategy not only avoids the shortcomings of Transformer in extracting edge detail features, but also prevents the loss of edge detail features and improves the quality of features, which helps to reconstruct high-quality fusion results.

3.4 Edge details guided feature fusion

In order to effectively use the edge detail features of the CNN branch to supplement the features in the Transformer branch in fusion image reconstruction, this paper proposes a method to synthesize the edge detail fusion results of the CNN branch and the fusion results of the Transformer branch to jointly construct the final fusion results. In order to effectively fuse the multimodal features extracted by TBFE. We design a multi-scale complementary mask fusion (MCMF) module to ensure its effectiveness. As shown in Figure 3, MCMF concatenate the features $\mathbf{F}_{ir}^i$ and $\mathbf{F}_{vi}^i$ from the output of the Transformer branch to obtain $\mathbf{F}_{ir+vi}^i$, which is feed into the convolutional layer to learn the weight map $\mathbf{M}^i$ for fusion. In this process, we apply three dilated convolutions with different dilation rates to the concatenated features, mining the importance information in different receptive fields in a

more flexible way. After concatenating three groups of results, the feature fusion is performed by 1 × 1 convolution, and the fusion weight map that reflects the importance of each position in the source image features is obtained through the Sigmoid activation function.

$$\mathbf{M}^i = \mathrm{Sigmoid}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{concat}\left(\mathrm{Conv}_{3\times3}\left(\mathbf{F}^i_{ir+vi}, \quad r=1\right),\right.\right.\right.$$
$$\left.\left.\left.\mathrm{Conv}_{3\times3}\left(\mathbf{F}^i_{ir+vi}, \quad r=2\right), \mathrm{Conv}_{3\times3}\left(\mathbf{F}^i_{ir+vi}, \quad r=3\right)\right)\right)\right) \tag{11}$$

where $\mathrm{Conv}_{3\times3}$ denotes 3 × 3 convolution and r is dilation rate.

The fusion feature can be expressed as:

$$\mathbf{F}^i = \mathbf{F}^i_{ir} \odot \mathbf{M}^i + \mathbf{F}^i_{vi} \odot \left(1 - \mathbf{M}^i\right) \tag{12}$$

where $\odot$ denotes hadamard product. Similar to the fusion of Transformer features, the edge detail features $\bar{\mathbf{F}}^i_{ir}$ and $\bar{\mathbf{F}}^i_{vi}$ extracted from the CNN branch are also fused in the above way to obtain the fusion result $\bar{\mathbf{F}}^3$ of the last RDB output features. The fused detailed features $\bar{\mathbf{F}}^3$ are concatenated with the fusion features of Transformer branch at three scales $\mathbf{F}^1$, $\mathbf{F}^2$, $\mathbf{F}^3$ to obtain $\mathbf{F}_f$. In order to ensure that both $\mathbf{F}_f$ and $\bar{\mathbf{F}}^3$ contain rich edge detail features, We reconstruct the edge detail feature maps by a reconstruction block (RB) for $\mathbf{F}_f$ and $\bar{\mathbf{F}}^3$ respectively, and make the reconstructed results consistent with the target feature maps. The RB used for reconstruction in this work consists of two 3 × 3 and one 1 × 1 convolutional layers, and the parameters are not shared between the different reconstruction block. Besides, this work uses the gradient detection operator to directly extract the gradient information of the source images, and fuse them to obtain high-quality target feature map $\mathbf{V}_{fus}$. The specific process is as follows:

$$\mathbf{V}_{fus}(i,j) = \begin{cases} \nabla\mathbf{X}_{ir}(i,j), if\,|\nabla\mathbf{X}_{ir}(i,j)| \geq \left|\nabla\mathbf{X}_{vi}(i,j)\right| \\ \nabla\mathbf{X}_{vi}(i,j), otherwise \end{cases} \tag{13}$$

where $\mathbf{V}_{fus}$ is the edge detail of the target image, ∇ is the Laplace operator, and (i, j) is the pixel coordinates.

4 Loss function

To ensure high visual quality fusion results, we use the L1 loss shown in Eq. 14 to optimize the parameters in the RDB:

$$\ell_f = \left\|\hat{\mathbf{Y}}^3 - \mathbf{V}_{fus}\right\|_1 \tag{14}$$

where $\hat{\mathbf{Y}}^3$ is the image after $\bar{\mathbf{F}}^3$ is reconstructed by RB. In the reconstruction process of the fusion results, the fusion image and the target feature map are reconstructed respectively with pixel loss, which is designed to limit the difference in intensity between real-world data and the reconstructed model result. $\mathbf{Y}_{fus}$ is the final fused image, and the reconstruction loss used in this work is as follows:

$$\ell_r = \|\mathbf{Y}_{fus} - \mathbf{X}_{ir}\|_1 + \|\mathbf{Y}_{fus} - \mathbf{X}_{vi}\|_1 + \left\|\hat{\mathbf{Y}}_f - \mathbf{V}_{fus}\right\|_1 \tag{15}$$

where $\hat{\mathbf{Y}}_f$ is the reconstructed image of $\mathbf{F}_f$ by RB. The total loss of the network is expressed as:

$$\ell_{total} = \ell_f + \ell_r \tag{16}$$

5 Experiments

5.1 Dataset

KAIST[1] and FLIR[2] are the two most commonly used datasets in the field of infrared and visible image fusion based on deep learning. Among them, there are 95,000 infrared and visible image pairs in the KAIST dataset and 14,452 image pairs in the FLIR dataset. In order to improve the generalization ability of the training model, 3,000 image pairs are randomly selected from the two datasets respectively, and a total of 6,000 image pairs from the training set of the proposed algorithm in this work. To verify the effectiveness of the method, 49 pairs of widely used infrared and visible images are randomly selected from the three datasets TNO[3], VOT2020-RGBT[4] and RoadScence[5] to construct the test set in this work. Among them, 39 pairs of images are from TNO and VOT2020-RGBT datasets and 10 pairs of images are from the RoadScence dataset. The test samples are shown in Figure 4.

5.2 Training details

In the training phase, each infrared and visible image pair is randomly cropped into 140 × 140 image blocks to achieve data enhancement. In this work, Adam [47] is used as the optimizer of the network, the training batchsize is set to 4, and a total of 30 epochs are iterated. The initial value of learning rate is set to 1×10^{-4}, and decays at the 5-th, 10-th, and 20-th epochs, respectively, with a decay rate of 0.5. The code of our method is implemented by using the PyTorch framework with NVIDIA GTX 3090, and the software environment is UBUNTU20.2, Python3.8 and PyTorch1.9.

5.3 Evaluation metrics

In order to objectively evaluate the fusion performance, six commonly used image fusion metrics are used in this work to assess the quality of fusion results from four perspectives. They are cross entropy (Q_{CE}) [48]; Entropy (Q_{EN}) [49]; gradient-based fusion performance ($Q_{AB/F}$) [50]; Chen -Blum metric (Q_{CB}) [51],; Chen-Varshney metric (Q_{CV}) [52]and Structural similarity index measure (Q_{SSIM}) [53]. Among them, Q_{CE} and Q_{EN} are metrics based on information theory ($Q_{AB/F}$) is a metrics based on image features, Q_{CB} and Q_{CV} are metrics based on human perception, and Q_{SSIM} is a metrics based on structural similarity of images. Among these six metrics, lower values for Q_{CE} and Q_{CV} indicate better quality of

1 https://soonminhwang.github.io/rgbt-ped-detection/.

2 https://www.flir.ca/oem/adas/adas-dataset-form/.

3 https://figshare.com/articles/dataset/TNO-Image-Fusion-Dataset/1008029.

4 https://www.votchallenge.net/vot2020/dataset.

5 https://github.com/hanna-xu/RoadScene.

FIGURE 4
Some test images from TNO, VOT2020-RGBT and RoadScence datasets.

TABLE 1 Quantitative evaluation of different fusion methods on 39 pairs of images from TNO and VOT2020-RGBT datasets. The red font indicates the optimal results and the blue font indicates the sub-optimal results.

Methods	Q_{CE} ↓	Q_{EN} ↑	$Q_{AB/F}$ ↑	Q_{CB} ↑	Q_{CV} ↓	Q_{SSIM} ↑
ADF	1.6913	6.4384	0.4021	0.4812	609.7199	1.4246
GTF	1.0113	6.7083	0.3346	0.4220	1,237.5143	1.4101
LatLRR	2.5355	6.7744	0.3596	0.4767	656.1416	1.1879
FusionGAN	2.2057	6.5336	0.2299	0.4045	1,025.8828	1.3489
SDNet	1.7178	6.6736	0.4527	0.4650	825.2304	1.4324
RFN	1.7270	6.8364	0.3602	0.4723	642.0348	1.4226
Ours	1.5498	6.9070	0.4815	0.5068	564.0687	1.4476

fusion results, while higher values for the other indicators indicate better fusion performance.

5.4 Comparison with state-of-the-arts

In order to verify the effectiveness of the proposed method, we compare our method with six advanced infrared and visible image fusion methods, including ADF [54], GTF [55], LatLRR [56], FusionGAN [23], SDNet [57], and RFN [22].

Figure 5 shows the fusion results of different methods on six groups of test images. In order to facilitate the fusion quality evaluation from the perspective of visual effect, we zoom in local area of the fusion results. It can be seen that the proposed method can not only preserve the salient information in the infrared image, but also maintain the edge detail information in the visible image. In detail, the outline of the infrared salient information is blurred, and

FIGURE 5
Fusion results of different methods on six pairs of images from TNO and VOT2020-RGBT datasets.

TABLE 2 Quantitative evaluation of different fusion methods on 10 pairs of images from RoadScence datasets. The red font indicates the optimal results and the blue font indicates the sub-optimal results.

Methods	Q_{CE} ↓	Q_{EN} ↑	$Q_{AB/F}$ ↑	Q_{CB} ↑	Q_{CV} ↓	Q_{SSIM} ↑
ADF	1.5202	6.7251	0.4084	0.4398	954.7853	1.3249
GTF	0.5876	7.0637	0.3708	0.4211	1964.4652	1.3212
LatLRR	2.0122	5.8881	0.3943	0.4285	880.5359	1.2419
FusionGAN	1.6016	6.9790	0.2939	0.4204	1,546.3186	1.2255
SDNet	1.6867	7.0493	0.4497	0.4604	1,587.5962	1.3285
RFN	1.3567	6.8826	0.3585	0.4390	1,398.5784	1.3139
Ours	1.5661	7.1463	0.4634	0.4839	741.7628	1.3601

the edge detail is not preserved enough in the fusion results obtained by FusionGAN, RNF, and ADF. In contrast, other methods effectively retain significant information, but the loss of spatial detail is more pronounced, as shown in the zoomed-in area of GTF fusion results. Similar phenomena can be observed in the remaining other fusion results. On the whole, the proposed method can more fully retain the significant edge detail information of the source images, and show better fusion performance. As for objective metrics, the proposed method has achieved excellent performance on Q_{EN}, $Q_{AB/F}$, Q_{CB}, Q_{CV} and Q_{SSIM}, as shown in Table 1, which further verifies the effectiveness of our method.

To further verify the effectiveness of the proposed method, we deploy the above comparison methods to the test data selected from the Roadscene. Figure 6 shows the fusion results of different

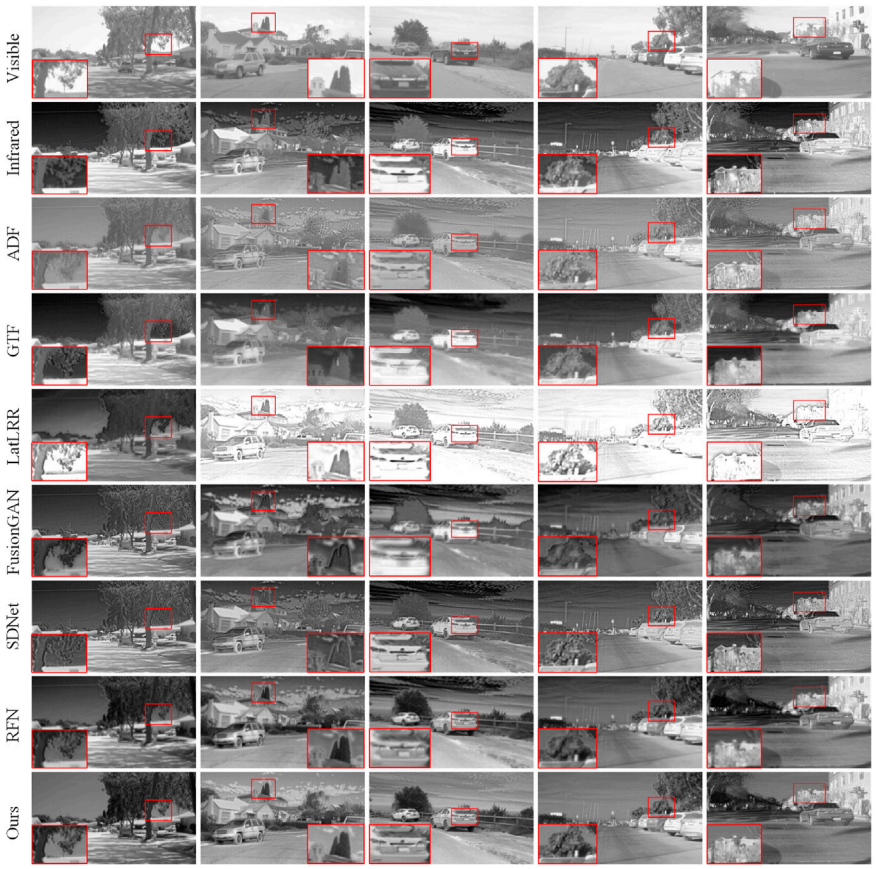

FIGURE 6
Fusion results of different methods on five pairs of images from RoadScence.

TABLE 3 Analysis of the effectiveness of different functional modules.

Methods	$Q_{CE}\downarrow$	$Q_{EN}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CB}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
w/o $\mathbf{H}_j$	1.5721	6.8866	0.4592	0.4857	600.7409	1.4166
w/o TFCB	1.5911	6.9014	0.4762	0.4780	569.5434	1.4244
w/o MCMF	1.5593	6.8715	0.4754	0.4812	588.4187	1.4406
w/o $\mathbf{V}_{fus}$	1.5557	6.8911	0.4547	0.4828	689.1718	1.4122
Ours	1.5498	6.9070	0.4815	0.5068	564.0687	1.4476

TABLE 4 The effect of different nuber of TFCBs on fusion performance.

Methods	$Q_{CE}\downarrow$	$Q_{EN}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CB}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
2 TFCBs	1.5687	6.8988	0.4805	0.4982	565.9757	1.4415
3 TFCBs	1.5498	6.9070	0.4815	0.5086	564.0687	1.4476
4 TFCBs	1.5425	6.8697	0.4805	0.4866	573.2399	1.4488

fusion methods on five pairs of test images from Roadscene. Deep learning-based fusion methods (FusionGAN, SDNet, and RFN) have better visual performance than traditional fusion methods (ADF, GTF, and LatLRR). Traditional fusion methods are limited by hand-designed fusion rules, resulting in problems such as too bright, too dark, or loss of detail information. FusionGAN adopts the adversarial learning network structure and lacks constraints on spatial consistency, causing blurred edge details in the obtained fusion results, which affects the visual quality improvement of the fusion images. SDNet, considers the spatial gradient information in network design, so it has certain advantages in detail retention, but its ability of remaining texture information is weak. Similar problem exists in RFN., in contrast, our fusion results have two advantages. First, the significant information of the infrared image can be well retained, so that the fusion results can further highlight the target, which is conducive to subsequent tasks (such as target detection, instance segmentation, etc.). Second, more texture detail information can be retained, ensuring the quality of the fusion results to a certain extent. In order to evaluate the quality of the fused images more comprehensively, we use six commonly used objective evaluation metrics to evaluate the quality of the fused images. From Table 2, results of the proposed method reach the optimal on five indicators of Q_{EN}, $Q_{AB/F}$, Q_{CB}, Q_{CV}, and Q_{SSIM}, which further proves the effectiveness and superiority of the proposed method.

5.5 Ablation study

In order to verify the influence of different components on the fusion performance of the proposed method, we apply ablation experiments on each module.

In the validation of the input of edge detail information, we use the source images instead of edge details as the input of CNN branch to verify the influence of edge details on fusion results (**w/o** $\mathbf{H}_j$). In the implantation of edge detail feature, we use two-stage feature compensation block (TFCB) to compensate the information of the Transformer branch. In order to verify the effectiveness of edge detail feature implantation, we add the features of two branch to replace TFCB module (**w/o** TFCB). In edge detail-guided feature fusion, the multi-scale complementary mask fusion (MCMF) module is the key. To verify its effectiveness, MCMF is replaced by conventional feature channel concatenation and 1 × 1 convolution to achieve feature fusion (**w/o** MCMF). The target edge detail map $\mathbf{V}_{fus}$ is used to enrich the features of the reconstructed fusion result with edge detail information. To demonstrate its validity, we directly removes it from the model (**w/o** $\mathbf{V}_{fus}$). The effectiveness of the above modules is tested on 39 pairs of images from TNO and VOT2020-RGBT. The effectiveness of the different components can be seen in Table 3.

5.6 Hyper-parameter analysis

The number of TFCBs determines the depth of the network, whose impact on the final performance can be seen in Table 4. When the number of TFCBs modules is three, the optimal result is obtained among the six evaluation indexes overall, so we set the number of TFCBs to three.

6 Conclusion

In order to effectively maintain the edge detail information of the source images, we propose a infrared and visible image fusion method with edge detail implantation, which adopts a two-branch feature representation framework. One branch is based on Transformer, which is mainly used to directly extract features from input source images. The other is CNN feature extraction branch, which is mainly used to extract image edge details features. Features extracted by CNN branch are implanted into the Transformer branch to alleviate the shortcomings of the Transformer branch in extracting edge detail features. In addition, so as to further ensure that the edge details of the source image can be effectively retained in the fusion results, a feature fusion method guided by edge details is proposed, which uses the fused edge detail features of CNN branch to guide the feature fusion of Transformer branch. A large number of experimental results prove the effectiveness of our method.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JL responsible for paper scheme design, experiment and paper writing. YZ guide the paper scheme design and revision. FL guide to do experiments and write papers. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 62161015). Name of Fund: Research on multi-source image fusion algorithm unconstrained by registration and resolution consistency.

Conflict of interest

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

Publisher's note

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

References

1. Han J, Bhanu B. Fusion of color and infrared video for moving human detection. *Pattern Recognition* (2007) 40:1771–84. doi:10.1016/j.patcog.2006.11.010

2. Singh R, Vatsa M, Noore A. Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition. *Pattern Recognition* (2008) 41:880–93. doi:10.1016/j.patcog.2007.06.022

3. Kumar P, Mittal A, Kumar P. Fusion of thermal infrared and visible spectrum video for robust surveillance. In: Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006; December 13-16, 2006; Madurai, India (2006). p. 528–39.

4. Tang L, Deng Y, Ma Y, Huang J, Ma J. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA J Automatica Sinica* (2022) 9:2121–37. doi:10.1109/jas.2022.106082

5. Liu Y, Shi Y, Mu F, Cheng J, Li C, Chen X. Multimodal mri volumetric data fusion with convolutional neural networks. *IEEE Trans Instrumentation Meas* (2022) 71:1–15. doi:10.1109/tim.2022.3184360

6. Liu Y, Wang Z. Dense sift for ghost-free multi-exposure fusion. *J Vis Commun Image Representation* (2015) 31:208–24. doi:10.1016/j.jvcir.2015.06.021

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

8. Li H, Yang M, Yu Z. Joint image fusion and super-resolution for enhanced visualization via semi-coupled discriminative dictionary learning and advantage embedding. *Neurocomputing* (2021) 422:62–84. doi:10.1016/j.neucom.2020.09.024

9. Vanmali AV, Gadre VM. Visible and nir image fusion using weight-map-guided laplacian–Gaussian pyramid for improving scene visibility. *Sādhanā* (2017) 42:1063–82. doi:10.1007/s12046-017-0673-1

10. Zheng M, Qi G, Zhu Z, Li Y, Wei H, Liu Y. Image dehazing by an artificial image fusion method based on adaptive structure decomposition. *IEEE Sensors J* (2020) 20: 8062–72. doi:10.1109/jsen.2020.2981719

11. Li H, Yu Z, Mao C. Fractional differential and variational method for image fusion and super-resolution. *Neurocomputing* (2016) 171:138–48. doi:10.1016/j.neucom.2015.06.035

12. Mitianoudis N, Antonopoulos SA, Stathaki T. Region-based ica image fusion using textural information. In: 2013 18th International Conference on Digital Signal Processing (DSP) (IEEE); Jul. 01 - 03, 2013; Greece (2013). p. 1–6.

13. Li H, He X, Tao D, Tang Y, Wang R. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognition* (2018) 79:130–46. doi:10.1016/j.patcog.2018.02.005

14. Zhang Q, Liu Y, Blum RS, Han J, Tao D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Inf Fusion* (2018) 40: 57–75. doi:10.1016/j.inffus.2017.05.006

15. Li H, He X, Yu Z, Luo J. Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior. *Inf Sci* (2020) 523:14–37. doi:10.1016/j.ins.2020.03.009

16. Xu H, Ma J, Jiang J, Guo X, U2fusion LH. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans Pattern Anal Machine Intelligence* (2020) 44:502–18. doi:10.1109/TPAMI.2020.3012548

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

18. Zhan L, Zhuang Y, Huang L. Infrared and visible images fusion method based on discrete wavelet transform. *J Comput* (2017) 28:57–71.

19. Yang B, Li S, Sun F. Image fusion using nonsubsampled contourlet transform. In: Fourth International Conference on Image and Graphics (ICIG 2007) (IEEE); 22-24 August 2007; Chengdu, China (2007). p. 719–24.

20. Li H, Qiu H, Yu Z, Zhang Y. Infrared and visible image fusion scheme based on nsct and low-level visual features. *Infrared Phys Tech* (2016) 76:174–84. doi:10.1016/j.infrared.2016.02.005

21. Wang X, Hua Z, Li J. Cross-UNet: Dual-branch infrared and visible image fusion framework based on cross-convolution and attention mechanism. *Vis Comput* (2022). doi:10.1007/s00371-022-02628-6

22. Li H, Wu XJ, Kittler J. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Inf Fusion* (2021) 73:72–86. doi:10.1016/j.inffus.2021.02.023

23. Ma J, Yu W, Liang P, Li C, Fusiongan JJ. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf fusion* (2019) 48:11–26. doi:10.1016/j.inffus.2018.09.004

24. Zhao Z, Xu S, Zhang C, Liu J, Li P, Zhang J. *Didfuse: Deep image decomposition for infrared and visible image fusion* (2020). *arXiv preprint arXiv:2003.09210*.

25. Liu Y, Liu S, Wang Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf fusion* (2015) 24:147–64. doi:10.1016/j.inffus.2014.09.004

26. Gao Y, Ma J, Yuille AL. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. *IEEE Trans Image Process* (2017) 26:2545–60. doi:10.1109/tip.2017.2675341

27. Li H, Wang Y, Yang Z, Wang R, Li X, Tao D. Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *IEEE Trans Instrumentation Meas* (2019) 69:1082–102. doi:10.1109/tim.2019.2912239

28. Zhu Z, Yin H, Chai Y, Li Y, Qi G. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inf Sci* (2018) 432:516–29. doi:10.1016/j.ins.2017.09.010

29. Yin M, Duan P, Liu W, Liang X. A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation. *Neurocomputing* (2017) 226:182–91. doi:10.1016/j.neucom.2016.11.051

30. Mo Y, Kang X, Duan P, Sun B, Li S. Attribute filter based infrared and visible image fusion. *Inf Fusion* (2021) 75:41–54. doi:10.1016/j.inffus.2021.04.005

31. Li H, Wu XJ. Densefuse: A fusion approach to infrared and visible images. *IEEE Trans Image Process* (2018) 28:2614–23. doi:10.1109/tip.2018.2887342

32. Li H, Zhao J, Li J, Yu Z, Lu G. Feature dynamic alignment and refinement for infrared-visible image fusion: Translation robust fusion. *Inf Fusion* (2023) 95:26.doi:10.1016/j.inffus.2023.02.011

33. Liu Y, Mu F, Shi Y, Chen X. Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion. *IEEE Signal Process. Lett* (2022) 29:1799–803. doi:10.1109/lsp.2022.3198594

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

35. Ma J, Zhao J, Jiang J, Zhou H, Guo X. Locality preserving matching. *Int J Comp Vis* (2019) 127:512–31. doi:10.1007/s11263-018-1117-z

36. Liu Y, Shi Y, Mu F, Cheng J, Chen X. Glioma segmentation-oriented multi-modal mr image fusion with adversarial learning. *IEEE/CAA J Automatica Sinica* (2022) 9: 1528–31. doi:10.1109/jas.2022.105770

37. Zhu Z, Liang H, Li Y, Qi G. A method for quality evaluation of multi-exposure fusion images with multi-scale gradient magnitude. In: Proceedings of 2021 Chinese Intelligent Systems Conference: Volume II; Nov 5-7, 2021; Zhanjiang, China (2022). p. 121–9.

38. Liu Y, Wang L, Li H, Chen X. Multi-focus image fusion with deep residual learning and focus property detection. *Inf Fusion* (2022) 86:1–16. doi:10.1016/j.inffus.2022.06.001

39. Ma J, Xu H, Jiang J, Mei X, Zhang XP. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans Image Process* (2020) 29:4980–95. doi:10.1109/tip.2020.2977573

40. Zhao Z, Xu S, Zhang J, Liang C, Zhang C, Liu J. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans Circuits Syst Video Tech* (2022) 32:1186–96. doi:10.1109/tcsvt.2021.3075745

41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30.

42. Ma J, Tang L, Fan F, Huang J, Mei X, Ma Y. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J Automatica Sinica* (2022) 9:1200–17. doi:10.1109/jas.2022.105686

43. Bandara WGC, Patel VM. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 19-25 June 2021 (2022) 1767–77.

44. Vs V, Valanarasu JMJ, Oza P, Patel VM. Image fusion transformer. In: 2022 IEEE International Conference on Image Processing (ICIP) (IEEE); October 16-19, 2022; Bordeaux France (2022). p. 3566–70.

45. Li J, Zhu J, Li C, Chen X, Yang B. Cgtf: Convolution-guided transformer for infrared and visible image fusion. *IEEE Trans Instrumentation Meas* (2022) 71:1–14. doi:10.1109/tim.2022.3175055

46. Park S, Vien AG, Lee C. Infrared and visible image fusion using bimodal transformers. In: 2022 IEEE International Conference on Image Processing (ICIP) (IEEE); October 16-19, 2022; Bordeaux France (2022). p. 1741–5.

47. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations; April 14-16, 2014; Banff, AB, Canada (2014). p. 109–19.

48. Bulanon D, Burks T, Alchanatis V. Image fusion of visible and thermal images for fruit detection. *Biosyst Eng* (2009) 103:12–22. doi:10.1016/j.biosystemseng.2009.02.009

49. Roberts JW, Van Aardt JA, Ahmed FB. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J Appl Remote Sensing* (2008) 2:023522. doi:10.1117/1.2945910

50. Xydeas CS, Petrovic V. Objective image fusion performance measure. *Elect Lett* (2000) 36:308–9. doi:10.1049/el:20000267

51. Chen Y, Blum RS. A new automated quality assessment algorithm for image fusion. *Image Vis Comput* (2009) 27:1421–32. doi:10.1016/j.imavis.2007.12.002

52. Chen H, Varshney PK. A human perception inspired quality metric for image fusion based on regional information. *Inf fusion* (2007) 8:193–207. doi:10.1016/j.inffus.2005.10.001

53. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. *IEEE Trans image Process* (2004) 13:600–12. doi:10.1109/tip.2003.819861

54. Bavirisetti DP, Dhuli R. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform. *IEEE Sensors J* (2015) 16:203–9. doi:10.1109/jsen.2015.2478655

55. Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf Fusion* (2016) 31:100–9. doi:10.1016/j.inffus.2016.02.001

56. Li H, Wu XJ. *Infrared and visible image fusion using latent low-rank representation* (2018). *arXiv preprint arXiv:1804.08992*.

57. Zhang H, Ma J. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int J Comp Vis* (2021) 129:2761–85. doi:10.1007/s11263-021-01501-8