Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(x): xxx–xxx
DOI:
Received: XX 2024
Revised: XX 2024
Accepted: XX 2024
Published: XX 2024

Type of article

CTFusion: CNN-Transformer-Based Self-Supervised Learning for Infrared and Visible Image Fusion

Keying Du[1,2], Liuyang Fang[1], Jie Chen[2], Dongdong Chen[3], and Hua Lai[2,*]

[1] Yunnan Key Laboratory of Digital Communications, Yunnan Communications Investment & Construction Group Co., LTD, Kunming, China

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

[3] CMA Meteorological Observation Centre, Beijing, China

* **Correspondence:** 405904235@qq.com (Hua Lai); Tel: +86-130-0865-6290.

Abstract: Infrared and visible image fusion (IVIF) is devoted to extracting and integrating useful complementary information from muti-modal source images. Current fusion methods usually require a large number of paired images to train the models in supervised or unsupervised way. In this paper, we propose CTFusion, a CNN-Transformer-based IVIF framework that uses self-supervised learning. The whole framework is based on an encoder-decoder network, where encoders are endowed with strong local and global dependencies modeling ability via th CNN-Transformer-based feature extraction (CTFE) module design. Thanks to the development of self-supervised learning, the model training does not require ground truth fusion images with simple pretext task. We design a mask reconstruction task according to the characteristics of IVIF, through which the network can learn the characteristics of both infrared and visible images and extract more generalized features. We evaluate our method and compared it to five competitive traditional and deep learning-based methods on three IVIF benchmark dataset. Extensive experimental results demonstrate that our CTFusion can achieve the best performance compared to the state-of-the-art methods in both subjective and objective evaluations.

Keywords: image fusion; self-supervised learning; Transformer

1. Introduction

Due to the hardware limitations of imaging, single sensor type or setup are often unable to fully represent imaging scenes [1, 2]. For example, visible images contain rich texture details, but susceptible to extreme environments and occlusion, leading to target loss in scenes. In contrast, infrared sensors are capable of imaging by capturing the thermal radiation information emitted by the objects, which effectively highlight pedestrians, vehicles, and other significant targets, but lack the detail description for the scenes [3]. In order to represent the scene accurately and effectively, image fusion is pushed forward to integrate the complementary features of multiple source views in the same scene, thus generating a high-quality image for the downstream high-level tasks or human perception [4]. Specifically, infrared and visible image fusion (IVIF) aims to integrate complementary information from the source images and generates a high-contrast fusion image that can both highlight salient objects and contain rich texture details [6]. In the early years, traditional methods usually use the related mathematical transformation and manual design of the fusion rules to realize the image fusion [3], such as wavelet [8], pyramid [9], and sparse representation [7]. However, these manually designed feature representation approaches and fusion rules poorly understand the inherent knowledge of images, which limits the ability to mine statistical characteristics of large samples and the generalizability.

Recently, deep learning has dominated the development of computer vision with its powerful feature extraction and expression capabilities. It has been used in all kinds of fields such as image classification [10], object detection [11, 44, 45], semantic segmentation [12] and image fusion [40]. In order to overcome the shortcomings of traditional algorithms, researchers explore a large number of image fusion methods based on deep learning, which can be divided into CNN-based and GAN-based infrared and visible image fusion frameworks. Nevertheless, these models exhibit a major weakness: the lack of ground-truth fused images. Some algorithms [20, 34] choose to generate ground-truth fusion result with existing state-of-the-arts (SOTA) fusion methods, whose fusion quality cannot be promised since it highly count on the quality of the produced "ground-truth". Moreover, most existing deep learning-based IVIF methods utilize convolutional neural networks (CNNs) for feature extraction, but CNNs fail to model longrange dependencies owing to their small receptive field, which is an inherent limitation [5].

In a word, there are two major problems in most existing IVIF task. On the one hand, the lack of ground-truth fused images. As mentioned above, though some methods try to generate so-called ground-truth fused images using other SOTA IVIF approaches, the quality of those produced ground-truths cannot be promised, thus largely affect the consequent fusion results. On the other hand, failure to capture long-range dependencies of CNNs due to their small receptive field becomes a weakness in most deep learning based IVIF methods.To address the aforementioned issues, we propose a CNN-Transformer-based infrared and visible image fusion framework utilizing self-supervised learning, dubbed CTFusion. Pretext task using masked image reconstruction helps better extract features of source images. CNN-Transformer-based encoder structure can utilize both local and global information. Inspired by [13], we adopt a specific image augmentation strategy that will mask some patches of the original images $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{\boldsymbol{vis}}$ with noise to generate two 'source images', $\widetilde{\boldsymbol{I}}_{ir}$ and $\widetilde{\boldsymbol{I}}_{vis}$. Afterwards, they are fed into the CNN-Transformer-based encoders to excavate the intrinsic features $\boldsymbol{f}_{ir}$ and $\boldsymbol{f}_{vis}$ in source images. We then apply two decoders $\boldsymbol{D}_{\boldsymbol{ir}}$ and $\boldsymbol{D}_{vis}$ to produce the repaired images of $\widetilde{\boldsymbol{I}}_{ir}$ and $\widetilde{\boldsymbol{I}}_{vis}$. In addition, a self-cross perceptual feature fusion (S-CPFF) strategy is elaborated, by which we combine features $\boldsymbol{f}_{ir}$ and $\boldsymbol{f}_{vis}$ together to generate the fusion result of $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$. The idea of our proposed method can be applied to common image fusion scenarios, since the above mentioned two problems are universal in image processing field.

In summary, the main contributions of this paper are summarized as follows:

- We propose a self-supervised IVIF framework by designing a mask reconstruction task which no longer needs ground-truth to better excavate intrinsic information lying in infrared and visible source images.

- To compensate for the defect in establishing long-range dependencies in CNN-based architectures, we design an encoder that combines a CNN-Block with a Transformer-Block, which enables the network to utilize both local and global information during feature extraction.
- S-CPFF module is devised to help enhance the extracted modality-specific and modality-common features, obtaining final fusion result with high quality.
- Extensive experiments conducted on three publicly available datasets demonstrate the effectiveness of our method, as well as show its superior performance with other state-of-the-art (SOTA) models.

2. Related works

2.1. Traditional image fusion methods

Traditional fusion frameworks usually realize image fusion in the transform domain and spatial domain through designing appropriate feature extraction details and fusion rules, which generally contain two major categories, including multi-scale transform-based [14, 38] and sparse representation-based [15, 39, 41, 42, 43].

Multi-scale transform-based methods first decompose source images into several levels, as the feature extraction, then fuse corresponding layers with particular rules, and reconstruct the target images accordingly, where popular transforms used for decomposition and reconstruction include wavelet [16], pyramid [17], curvelet [18], and their revised versions. However, these methods typically tend to miss out image details in the fused results and lead to halos or undesirable artifacts in the fused result due to the fixed bases used in the multi-scale transform-based methods. The key of sparse representation-based methods is to build over complete dictionaries from a large number of natural images to possibly represent the source images with linear combination of sparse bases. Although SR-based methods have achieved promising performance, a limited number of dictionaries cannot reflect the full information of input images, obscuring details such as edges and textures in the source images.

2.2. Deep learning-based image fusion methods

In deep learning-based algorithms, two source images from different modalities are directly input into a fusion network and then the network outputs the fused image. Specifically, Liu et al. [19] proposed a method based on convolutional neural networks, which can deal with activity level measurement and weight assignment in infrared and visible image fusion as a whole to overcome the difficulty of manual design. To get more useful features from source images, Liu et al. [20] presented a novel encoding network combined with convolutional layers, a fusion layer, and dense block in which the output of each layer is connected to every other layer. With the development of generative adversarial network (GAN), more and more GAN-based IVIF methods boomed out. Although CNN has made great achievements in the field of supervised learning, it still has not progressed much in the unsupervised learning. In order to fill the gap between supervised learning and unsupervised learning of CNN, Ma et al. [21] established an adversarial game between a generator and a discriminator, which enabled that the final fused image simultaneously kept the thermal radiation in an infrared image and the textures in a visible image. Meanwhile, generic image fusion frameworks also achieved surprising performance. Li et al. [36] proposed a meta learning-based deep framework for the fusion of infrared and visible images which can accept the source images of different resolutions and generate the fused image of arbitrary resolution just with a single learned model. Zhang et al. [22] proposed a squeeze-and-decomposition network named SDNet to realize multi-modal and digital photography image fusion in real time. Xu et al. [23] used feature extraction and information measurement to automatically estimate the importance of corresponding source images and came up with adaptive information preservation degrees, solving different fusion problems. Recently, Tang et al. [24] and Liu et al. [25] bridged the gap between image fusion and high-level vision tasks, facilitating the high-level vision tasks with the proposed frameworks.

Though existing approaches devised complicated fusion rules and loss functions to achieve effective fusion, they still fail to effectively learn the characteristics of infrared and visible images for not devising specific task to explore intrinsic features in source images. It is a consensus that feature extraction is a pivotal step in image fusion. If the extracted features cannot represent rich while comprehensive characteristics of source images, qualities of fusion results will be definitely declined. In contrast, we design self-supervised mask reconstruction task to deeply excavate the intrinsic characteristics of infrared and visible images so that our network is able to achieve high-quality IVIF fusion.

2.3. Vision Transformer

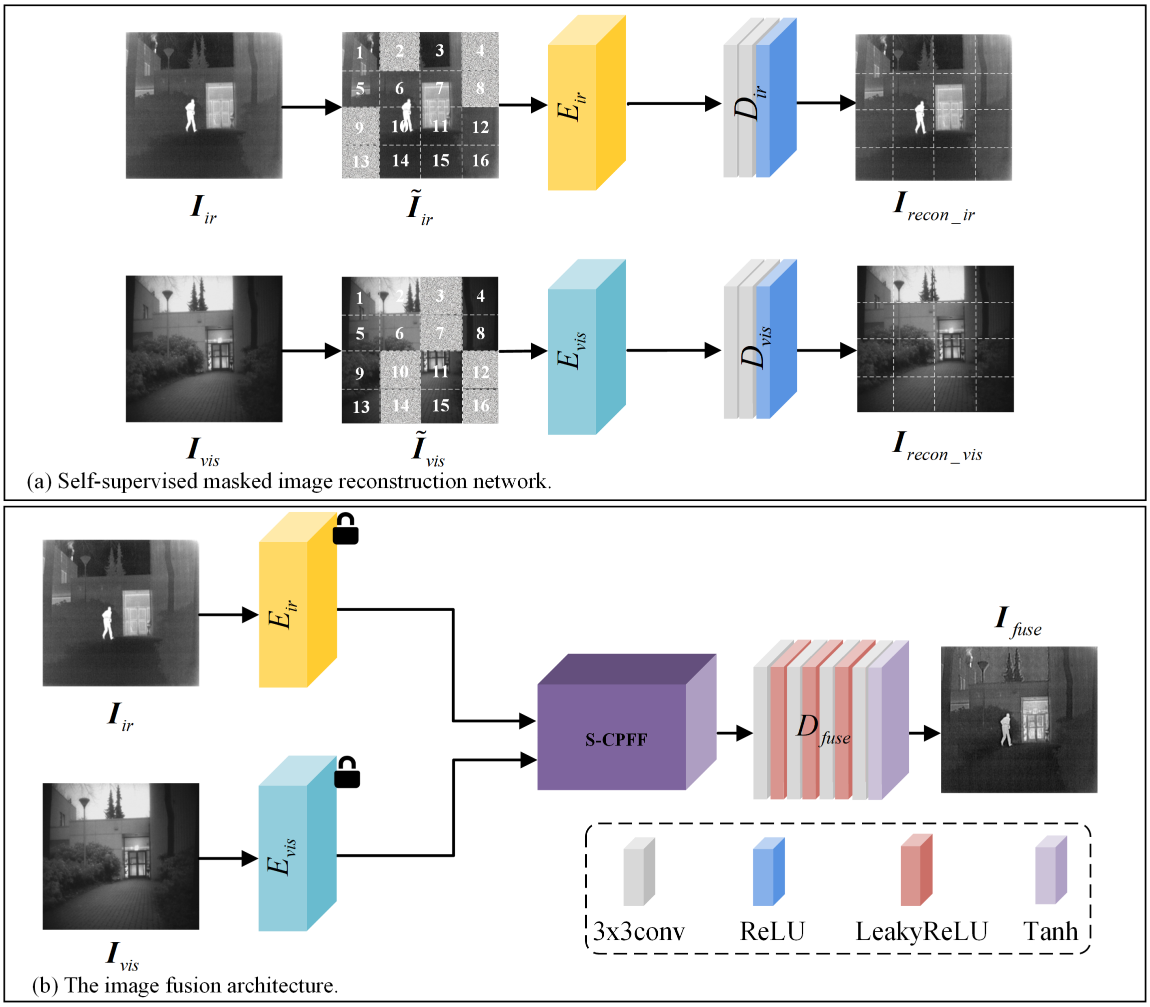
Transformer was first proposed by Vaswani et al. [46] for machine translation. Its ability to extract features from the global level and effectively depict the correlation between features at different locations has attracted wild attention in the community. Later, researchers made great success in introducing Transformer into computer vision tasks, such as image processing [47], object detection [48], semantic segmentation [49], etc. Dosovitskiy et al. [50] introduced Transformer to image classification task for the first time, proposing Vision Transformer (ViT). Based on ViT, a series of ViT variants were proposed to improve their performance [51, 52]. Particularly, in the field of image fusion, Vibashan et al. [53] proposed a Transformer based infrared and visible image fusion method. This method used Transformer's encoder to extract image features, obtained the fused features with Spatial-Transformer, and finally reconstructed the fused image through Transformer's decoder.

Since Transformer has a stronger ability to model long-range dependencies, it is suitable to extract global image features. In contrast, CNN is apt to capture local image features and describe low-level visual features such as structure and texture details for it extracts image features through convolutional kernels, whose receptive field is limited. To integrate advantages of them, we design an encoder that combines a CNN-Block with a Transformer-Block, which enables the network to utilize both local and global information during feature extraction.

3. Method

3.1. Overview

As mentioned above, lack of ground-truth fused images remains a tricky problem in IVIF tasks. Therefore, we propose CTFusion to achieve fusion in a self-supervised way, where a pretext reconstruction task aiming at image understanding is elaborated. More specifically, a CNN-Transformer-based encoder is devised to compensate for the defect in establishing long-range dependencies in CNN-based architectures. After the parameters in encoders are optimized, we fix them during the fusion phase and then fuse the extracted features maps through tailored fusion net S-CPFF, to generate the fused image.



(a) Self-supervised masked image reconstruction network.

(b) The image fusion architecture.

Figure 1. Overall framework of our proposed method. (a) The proposed self-supervised masked image reconstruction network. (b) The image fusion architecture.

Pipeline of the encoder training process is shown in Figure 1(a), where we use the network to perform self-supervised image reconstruction task to enable encoders to extract intrinsic features lying in source images, i.e., to reconstruct the original image from the masked input image. Concretely, given an original image $\boldsymbol{I}_{in(in\in\{ir,vis\})} \in \mathbb{R}^{H\times W\times 3}$, masked image $\widetilde{\boldsymbol{I}}_{in}$ is generated by masking several non-overlapping patches with noise. Then, we feed the masked images into respective encoder $\boldsymbol{E}_{ir}$ and $\boldsymbol{E}_{vis}$ to obtain corresponding embeddings, which consists of a CNN-Transformer feature extraction (CTFE) module and a feature enhancement (FE) module. CTFEBlock integrates the advantages of CNN and Transformer to model both global and local dependencies. FE aggregates and enhances the features extracted from the CNN-Block and the Transformer-Block. Finally, the image features extracted by the encoders are sent to respective decoder $\boldsymbol{D}_{ir}$ and $\boldsymbol{D}_{vis}$ to reconstruct image $\boldsymbol{I}_{recon\_ir}$, $\boldsymbol{I}_{recon\_vis} \in \mathbb{R}^{H\times W\times 3}$.

After training the encoders, we then use them for image fusion with their parameters fixed, as shown in Figure 1(b). Specifically, two source images $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$ are first input to the trained encoders $\boldsymbol{E}_{ir}$ and $\boldsymbol{E}_{vis}$ to extract features, and then the fusion result is obtained by fusing the extracted features using the well-designed fusion network S-CPFF.

3.2. *Self-supervised mask reconstruction task*

In general, the goal of image fusion is to integrate complementary information from different source images into a synthetic image. Moreover, feature dependency excavation is also the key in image fusion, since the relation understanding is important in feature extraction. Features with rich semantic and structural information are ideal to obtain high-quality fusion results. We divide the input image into non-overlapping patches and then use a random mask M and Gaussian noise n to force encoders to excavate intrinsic information lying in source images.

$$\widetilde{\boldsymbol{I}}_{in} = M(\boldsymbol{I}_{in}) + \overline{M}(n)(in \in \{ir, vis\}), \tag{3.1}$$

where $\overline{M}(\cdot)$ is the logical negation operator of mask M.

For each source image paired $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$, they share parts of the scene information with the other while retaining some unique information. By randomly masking and filling the remaining with random noise, the encoders are forced to extract more information in source images, better understanding the relation between pixels. After pre-training, the encoders are able to extract more comprehensive features, which can be directly used for the following image fusion task.

3.3. *CNN-Transformer-based encoder-decoder framework*

Given source images $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$, we first randomly mask subregions and fill the remaining with noise to form $\widetilde{\boldsymbol{I}}_{ir}$ and $\widetilde{\boldsymbol{I}}_{vis}$ that will be sent to the CNN-Transformer-based encoders $\boldsymbol{E}_{ir}$ and $\boldsymbol{E}_{vis}$. Each encoder contains a CTFE module, and a FE module, whose detailed architectures are shown in Figure 2.

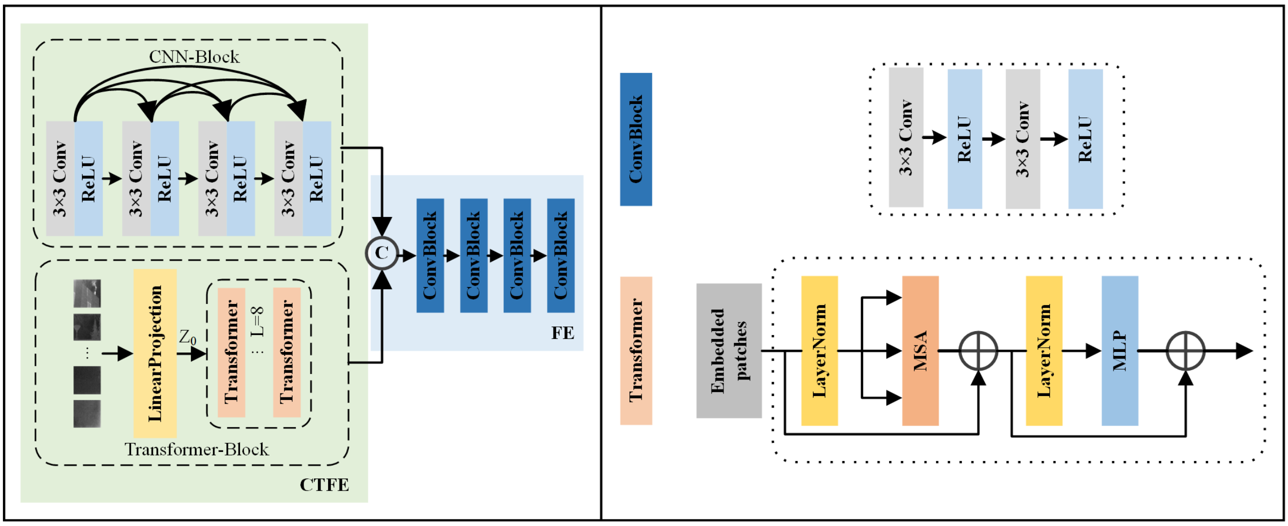


Figure 2. Detailed structures of CNN-Transformer-based encoder, ConvBlock and Transformer.

Given that CNN is adept at modeling local dependencies in images, while Transformer specializes in modeling global dependencies, we propose CTFE that combines the CNN and Transformer architecture to model both local and global dependencies in images. Specifically, the CNN-Block contains a residual dense block following the residual dense network [26]. As for the Transformer-Block, the masked image $\widetilde{\boldsymbol{I}}_{in(in\in\{ir,vis\})} \in \mathbb{R}^{H\times W\times 3}$ is firstly divided into a total of N patches with size $\frac{H}{P} \times \frac{W}{P}$, where $N = \frac{HW}{P^2}$ and P is the size of the patches. Passing the patches through a patch embeddings linear projection and L Transformer layers, we can obtain Transformer-embedded feature $\boldsymbol{f}^{tf}_{in(in\in\{ir,vis\})}$. Figure 2 illustrates the architecture of one Transformer layer, which consists of a multi-head attention (MSA) block and a multi-layer perceptron (MLP) block, where layer normalization (LN) is applied before every block and residual connections are applied after every block. The MLP block consists of two linear layers with a GELU activation function. In order to better integrate the local and global features extracted by CNN and Transformer blocks, we devise the feature enhancement module FE to aggregate and enhance the feature maps $\boldsymbol{f}^{cnn}_{in(in\in\{ir,vis\})}$ and $\boldsymbol{f}^{tf}_{in(in\in\{ir,vis\})}$. Concretely, we concatenate the two feature maps from the CNN-Block and the Transformer-Block in CTFE and send them into four sequentially connected ConvBlock layers to achieve feature enhancement, as shown in Figure 2.

$$\boldsymbol{f}^{en}_{in} = (ConvBlock([\boldsymbol{f}^{cnn}_{in}, \boldsymbol{f}^{tf}_{in}]))_{\times 4}, in \in \{ir, vis\}, \tag{3.2}$$

where each ConvBlock consists of two convolutional layers with a kernel size of 3 × 3, a padding of 1 and two ReLU activation layers, $[\cdot]$ denotes channel-wise concatenation.

We then feed the obtained feature maps $\boldsymbol{f}^{en}_{ir}$ and $\boldsymbol{f}^{en}_{vis}$ to decoders $\boldsymbol{D}_{ir}$ and $\boldsymbol{D}_{vis}$ each in which composed of two convolutional layers with a kernel size of 3 × 3, a padding of 1 and one ReLU activation layer to reconstruct the corresponding image.

In the mask reconstruction task, we encourage the network to not only learn the pixel-level image reconstruction, but also capture the structural and gradient information in the image. The loss of the reconstruction task in each branch can be formalized as follows:

$$\ell_{reconstruction} = \ell_{pixel} + \lambda_1 \ell_{structure} + \lambda_2 \ell_{TV}, \tag{3.3}$$

where ℓ_{pixel} is the L1 loss function, $\ell_{structure}$ is the structural similarity (SSIM) loss function, and ℓ_{TV} is the total variation loss function. λ_1 and λ_2 are two hyperparameters empirically set to 20.

ℓ_{pixel} ensures pixel-level reconstruction

$$\ell_{pixel} = \boldsymbol{I}_{recon\_in} - \boldsymbol{I}_{in}, in \in \{ir, vis\}, \tag{3.4}$$

where $\boldsymbol{I}_{recon\_in}$ is the output reconstructed image, and $\boldsymbol{I}_{in}$ represents the input unmasked source image.

To better help the model learn structural information from images, we use the structure loss as:

$$\ell_{structure} = 1 - SSIM(\boldsymbol{I}_{recon\_in}, \boldsymbol{I}_{in}), in \in \{ir, vis\}. \tag{3.5}$$

Furthermore, ℓ_{TV} in VIFNet [27] is used to facilitate gradient preservation in the source images and eliminate noise. It is formulated as follows:

$$\ell_{TV} = \sum_{x,y} \|R(x, y+1) - R(x, y)\|_2 + \|R(x+1, y) - R(x, y)\|_2. \tag{3.6}$$

where $R(x, y) = \boldsymbol{I}_{recon\_in}(x, y) - \boldsymbol{I}_{in}(x, y)$ $(in \in \{ir, vis\})$ denotes the difference between the input image and the reconstructed image, $\|\cdot\|_2$ is the L2 norm, and x, y represent the horizontal and vertical coordinates of the image's pixels, respectively.

3.4. *Self-cross perceptual feature fusion*

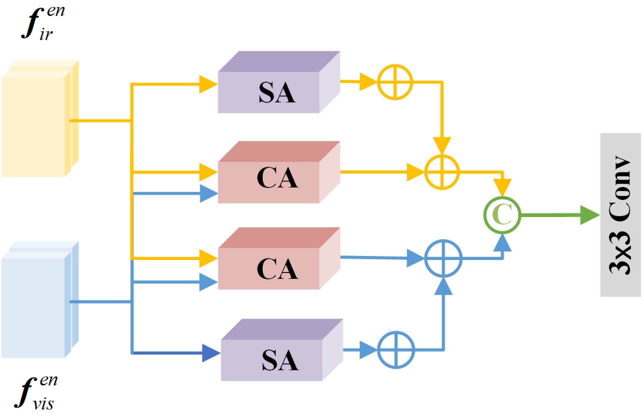


Figure 3. Detailed framework of S-CPFF.

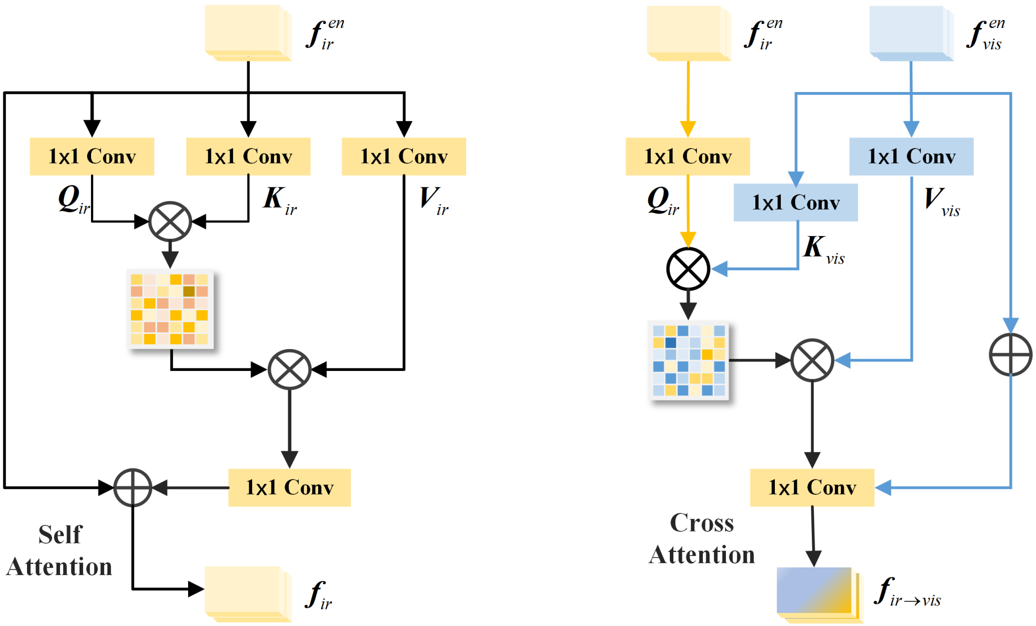


Figure 4. Detailed structures of SA and CA.

How to effectively fuse feature maps to obtain final fusion results remains challenging. Here we propose S-CPFF, which is able to highlight both self and mutual interested parts in feature maps, naturally improving the qualities of fused images. Concretely, the S-CPFF covers two self-attention (SA) and two cross-attention (CA). Mathematically, the SA process is denoted as:

$$\begin{aligned} f_{ir} &= \mathrm{softmax}(\frac{\mathbf{Q}_{ir}(\mathbf{K}_{ir})^T}{\sqrt{d}})\mathbf{V}_{ir}, \\ f_{vis} &= \mathrm{softmax}(\frac{\mathbf{Q}_{vis}(\mathbf{K}_{vis})^T}{\sqrt{d}})\mathbf{V}_{vis}, \end{aligned} \tag{3.7}$$

where $\mathbf{Q}_{ir}$, $\mathbf{Q}_{vis} \in \mathbb{R}^{H\times W\times C}$, $\mathbf{K}_{ir}$, $\mathbf{K}_{vis} \in \mathbb{R}^{H\times W\times C}$, and $\mathbf{V}_{ir}$, $\mathbf{V}_{vis} \in \mathbb{R}^{H\times W\times C}$ are the results of $\boldsymbol{f}_{ir}^{en}$ and $\boldsymbol{f}_{vis}^{en}$ passed through 1×1 convolution, respectively. $\sqrt{d}$ is a normalization factor, and T is transpose operation. The CA process is denoted as:

$$\begin{aligned} \boldsymbol{f}_{ir\rightarrow vis} &= \text{softmax}(\frac{\mathbf{Q}_{vis}(\mathbf{K}_{ir})^T}{\sqrt{d}})\mathbf{V}_{ir}, \\ \boldsymbol{f}_{ir\rightarrow vis} &= \text{softmax}(\frac{\mathbf{Q}_{ir}(\mathbf{K}_{vis})^T}{\sqrt{d}})\mathbf{V}_{vis}, \end{aligned} \tag{3.8}$$

where $vis \rightarrow ir$ denotes information flow from visible modal to infrared modal. Then we concatenate the results after respective SA and CA module to obtain the self-cross perceptual feature:

$$\begin{aligned} \boldsymbol{f}_{ir}' &= [\boldsymbol{f}_{ir}, \boldsymbol{f}_{vis\rightarrow ir}], \\ \boldsymbol{f}_{vis}' &= [\boldsymbol{f}_{vis}, \boldsymbol{f}_{ir\rightarrow vis}]. \end{aligned} \tag{3.9}$$

Please note that in the fusion phase, we fix the parameters of encoders $\boldsymbol{E}_{ir}$ adn $\boldsymbol{E}_{vis}$ learned in mask reconstruction task. We then concatenate $\boldsymbol{f}_{ir}'$ and $\boldsymbol{f}_{vis}'$, and feed the concatenated result to a 3×3 convolution to integrate all the features. Then the integrated feature is sent to the decoder $\boldsymbol{D}_{fuse}$ to attain the fused image:

$$\boldsymbol{I}_{fuse} = \boldsymbol{D}_{fuse}(Conv_{3\times 3}([\boldsymbol{f}_{ir}', \boldsymbol{f}_{vis}'])), \tag{3.10}$$

where the detailed structure of $\boldsymbol{D}_{fuse}$ is shown in Figure 1.

To retain rich edge and texture information in the fused image, we adopt joint gradient loss ℓ_{JGrad}, which is formulated as:

$$\ell_{JGrad} = \|\mathrm{O}(\max(|\nabla \boldsymbol{I}_{ir}|, |\nabla \boldsymbol{I}_{vis}|)) - \nabla \boldsymbol{I}_{fused}\|_1, \tag{3.11}$$

where ∇ is Laplacian gradient operator. $\max(\cdot)$ denotes taking the maximum value. $\mathrm{O}(|x|) = x$ denotes finding the original gradient value before taking its absolute value.

Besides, we also introduce the intensity loss to preserve the saliency targets in two input images, which can be expressed as:

$$\begin{aligned} \boldsymbol{\omega}_{ir} &= \boldsymbol{S}_{I_{ir}}/(\boldsymbol{S}_{I_{ir}} - \boldsymbol{S}_{I_{vis}}), \boldsymbol{\omega}_{vis} = 1 - \boldsymbol{\omega}_{ir}, \\ \ell_{int} &= \|(\boldsymbol{\omega}_{ir} \odot \boldsymbol{I}_{ir} + \boldsymbol{\omega}_{vis} \odot \boldsymbol{I}_{vis}) - \boldsymbol{I}_{fused}\|_1, \end{aligned} \tag{3.12}$$

where $\boldsymbol{S}_{I_{ir}}$ and $\boldsymbol{S}_{I_{vis}}$ denote saliency matrices of $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$, which can be computed according to references [28]. $\boldsymbol{\omega}_{ir}$ and $\boldsymbol{\omega}_{vis}$ are the weight maps for $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$, respectively. $\odot$ represents element-wise multiplying operation.

The overall fusion loss is computed by

$$\ell_{fuse} = \ell_{int} + \lambda_{JG}\ell_{JGrad}, \tag{3.13}$$

where λ_{JG} is the hyper-parameter set to 20.

4. Experiments and results

In this section, we evaluate the qualitative and quantitative performance of our proposed method by comparing to five SOTA methods including IFCNN [29], PMGI [30], CrossFuse [31],RFN-Nest [32] and FusionGAN [21]. Besides, we also implement abundant ablation studies to validate the effectiveness of the proposed modules.

4.1. Experimental configurations

Dataset: 300 multi-modality images from the M3FD [25] benchmark are selected and cropped to 360k patches with 256×256 pixels by random cropping and augmented as the training set in this paper. M3FD is a multi-modal dataset with multiple scenarios, where 4,200 aligned image pairs are divided into four typical types, i.e., Day, Cloudy, Night, and Challenge. We perform qualitative and quantitative experiments on three datasets (i.e., Roadscene, TNO, and MSRS). RoadScene is a wildly-used dataset for cross-modality image fusion. TNO dataset contains multi-spectral nighttime imagery of various military-relevant scenarios in grayscale. MSRS dataset contains 1,444 pairs of aligned infrared and visible images with high quality.

Evaluation metrics: For quantitative evaluation, five statistical metrics are selected to objectively assess the fusion performance, including correlation coefficient (CC) [56], cross entropy (CE), Q^{CV} [57], the sum of correlations of differences (SCD) [58], and structural similarity (SSIM) [59]. CC evaluates the degree of linear correlation between the fused image and source images. CE reflects the difference of grayscale information between fusion image and source images. The smaller the CE value, the smaller the difference between images, which indicates better fusion quality. Q^{CV} uses the Sobel operator to extract the edge information of the source images and the fusion result to obtain the edge intensity map G. Smaller Q^{CV} implies more in line with human visual perception. SCD reflects the correlation level between information transmitted to the fused image and corresponding source images. SSIM approximates image distortion. In addition, a fusion algorithm with larger CC, SCD, and SSIM indicates better fusion performance.

Implement details: The Adam optimizer [33]($\beta_1 = 0.9$, and $\beta_2 = 0.999$) is responsible for updating the network parameters with initial learning rate of 0.001, which decreases to 10^{-4} after 100 epochs. The epochs of self-supervised mask reconstruction task and training of S-CPFF are both set to 300 with batch size of 4. Our framework is implemented on PyTorch with an NVIDIA 3090 GPU. Please note that source images in all the above mentioned datasets are converted to gray to achieve fusion.

4.2. Comparative experiment

4.2.1. Experiments on RoadScene dataset

Qualitative results on the RoadScene benchmark are reported in Figure 5, where we highlight two regions in each example. As can be seen, the RFN-Nest and FusionGAN suffer blurred edges and background, and the IFCNN, PMGI, and CrossFuse all lose texture details to some extent. Instead, our method reserves the best image contrast and clear structure information.

Quantitative comparisons are shown in Table 1 where we use five metrics, i.e., CC, CE, Q^{CV}, SCD and SSIM to evaluate all comparison methods. Our method ranks first on the CC, CE, SCD and SSIM, indicating that the generated fused results are of higher similarity to the source images. For Q^{CV}, our method also achieves comparable results, which implies that the fusion images of our method are more real. Since our model utilizes Laplacian gradient operator to detect the edges of images, while Q^{CV} uses Sobel gradient operator, it might be the reason why our method performs suboptimally in Q^{CV}.

4.2.2. Experiments on TNO dataset

We select five pairs of infrared and visible images to visually observe the fusion performance of different algorithms on the TNO dataset. The visualized results are shown in Figure 6. As shown in the

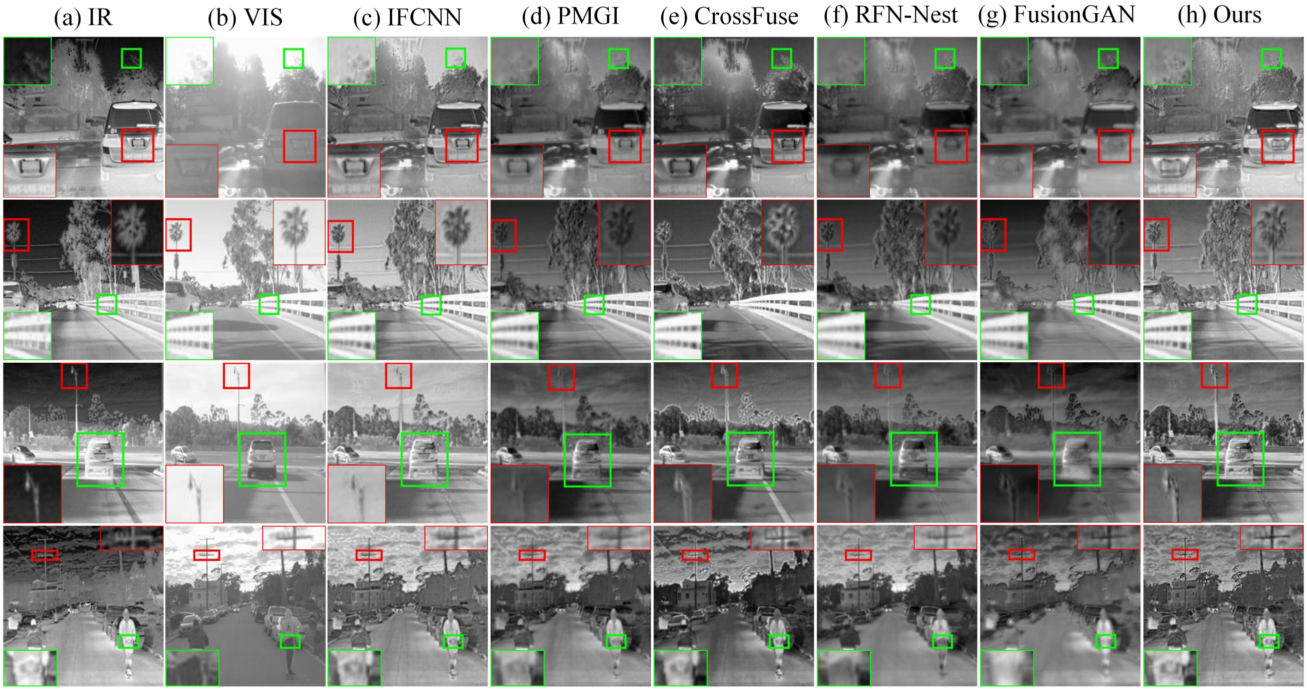


Figure 5. Vision quality comparison of our method with five SOTA fusion methods on the RoadScene dataset.

Table 1. Quantitative results of five SOTA methods and ours on 50 image pairs from RoadScene [23] dataset. Bold: best. Italic: second best.

Methods	CC	CE	Q^{CV}	SCD	SSIM
IFCNN[29]	*0.622*	0.976	**589.5**	1.245	*0.693*
PMGI[30]	0.596	1.328	1019.6	1.218	0.644
CrossFuse[31]	0.614	1.397	943.3	*1.397*	0.687
RFN-Nest[32]	0.582	*0.922*	983.2	1.373	0.603
FusionGAN[21]	0.577	2.308	1371.2	0.889	0.615
Ours	**0.641**	**0.763**	*820.9*	**1.447**	**0.702**

Table 2. Quantitative results of five SOTA methods and ours on 30 image pairs from TNO [35] dataset. Bold: best. Italic: second best.

Methods	CC	CE	Q^{CV}	SCD	SSIM
IFCNN[29]	0.643	1.734	**392.5**	1.215	*0.695*
PMGI[30]	0.651	1.781	496.1	1.222	0.676
CrossFuse[31]	*0.669*	*1.694*	943.3	1.349	0.682
RFN-Nest[32]	0.622	1.792	533.6	1.361	0.689
FusionGAN[21]	0.554	2.380	968.7	*1.451*	0.628
Ours	**0.675**	**1.499**	*433.6*	**1.476**	**0.701**

first column of Figure 6, our method keeps the best image contrast. It can be seen from the zoomed-in areas that our methods combine complementary as well as modality-common information in source images to the most extent. Meanwhile, the edges of salient targets from infrared images are clear and texture details from visible images are well kept in our fusion results.

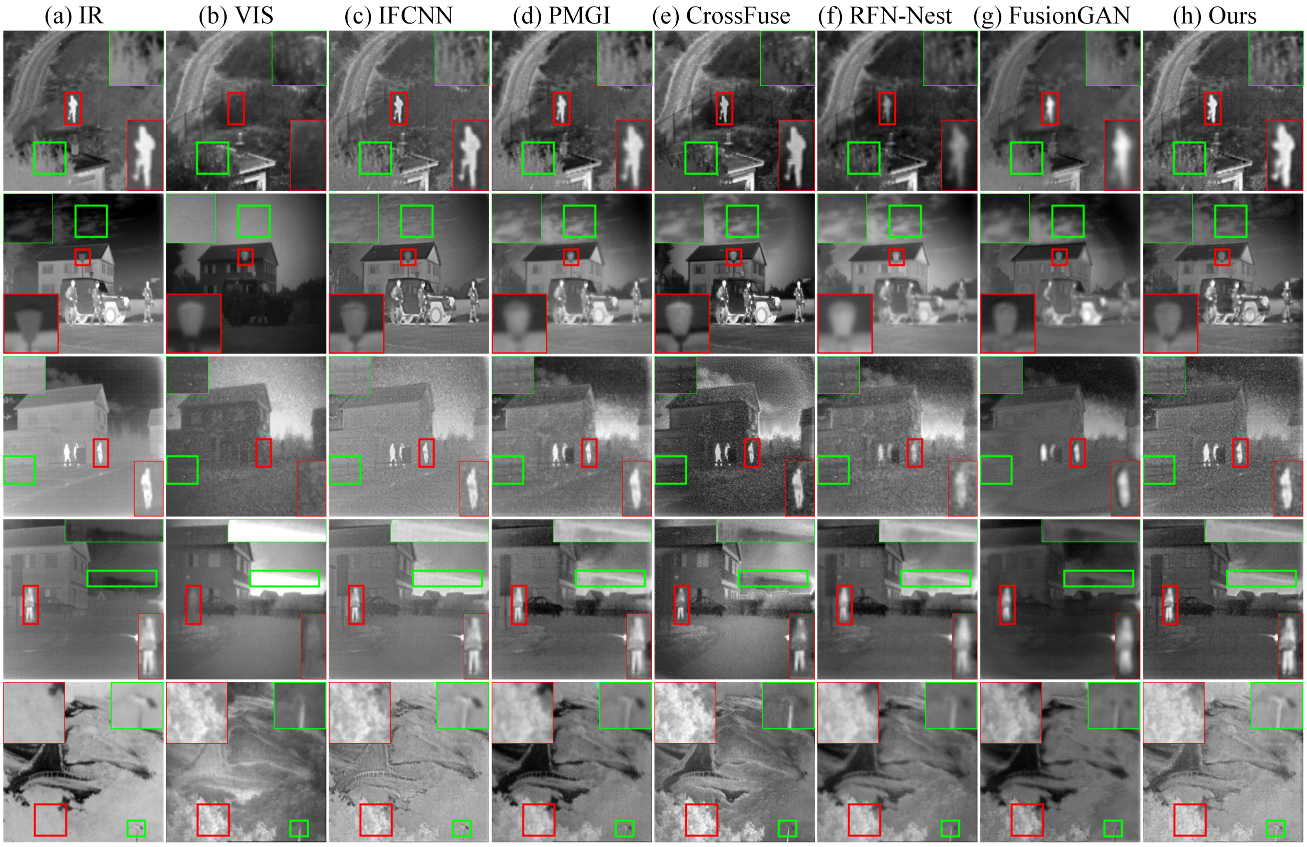


Figure 6. Vision quality comparison of our method with five SOTA fusion methods on the TNO dataset.

For example, in the second row of Figure 6, though the image contrast of CrossFuse is better than ours, it still lose clouds information. In the first, third rows, people in the foreground are bright in our fused results, while FusionGAN has blurred target edges.

Quantitative performance of our method on the TNO dataset show similar performances to those of

RoadScene.

4.2.3. Experiments on MSRS dataset

So as to visually evaluate the fusion performance of different algorithms on MSRS dataset, three pairs of infrared and visible images are selected, depicted in Figure 7. As illustrated in the red and green boxes in the image, our proposed method maintain favourable textures of source images while keeping the clearest salient edges.

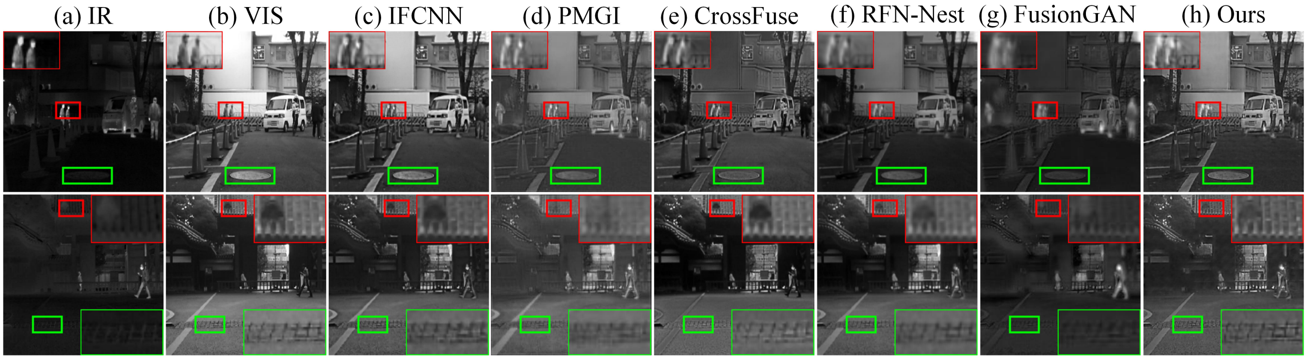


Figure 7. Vision quality comparison of our method with five SOTA fusion methods on the MSRS dataset.

We conduct quantitative comparisons on 40 image pairs from MSRS dataset to verify the effectiveness of our method, which is presented in Table 3. It can be seen that our method ranks first in four metrics and second in the CC metric. The CE, Q^{CV}, SCD and SSIM metrics demonstrate that our results contain more realistic information. As for CC, it directly matches images by their intensity, without using any analysis of the image structure. Hence CC is sensitive to intensity changes in the image. In general, image noise, changes in lighting intensity during imaging, and the use of different imaging equipment all cause changes in image intensity, which will further affect CC.

Table 3. Quantitative results of five SOTA methods and ours on 40 image pairs from MSRS [24] dataset. Bold: best. Italic: second best.

Methods	CC	CE	Q^{CV}	SCD	SSIM
IFCNN[29]	**0.551**	0.936	873.9	1.219	0.657
PMGI[30]	0.488	1.127	1292.3	*1.336*	0.628
CrossFuse[31]	0.527	1.248	865.7	1.245	*0.663*
RFN-Nest[32]	0.495	*0.891*	*823.2*	1.328	0.614
FusionGAN[21]	0.463	2.185	1587.8	0.713	0.605
Ours	*0.537*	**0.766**	**743.5**	**1.423**	**0.675**

In conclusion, our method is fully capable of excavating inherent important features in source images and integrating them into fused images. Thereby, our method is superior to other SOTA approaches and obtains high-quality fused images.

4.3. Analysis of generalization ability

To validate the generalization ability of our method, we conduct experiments on datasets for other image fusion tasks, including LLVIP [54] for color image fusion and CT-MRI [55] for medical image fusion. Fusion results are shown in Figure 8. From the qualitative results we can see that our proposed model perfectly complete other fusion tasks, which strongly proves the generalization ability of our method.

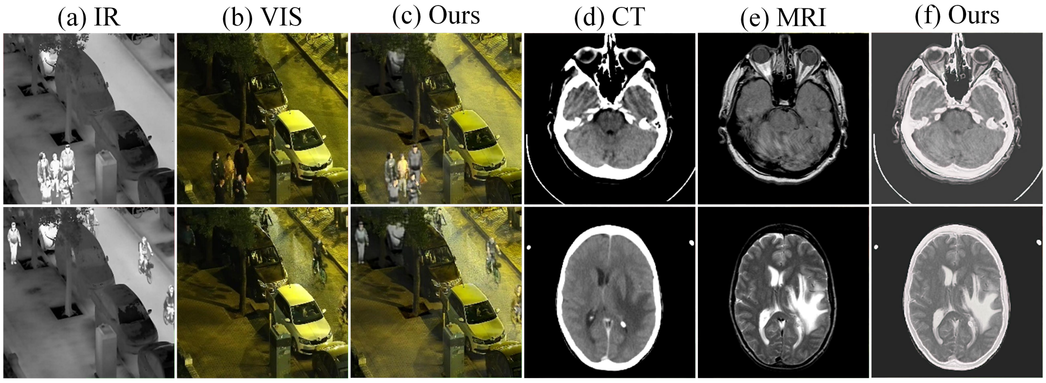


Figure 8. Vision effect of our method on the LLVIP and CT-MRI datasets. (a)-(c) are our fusion results on LLVIP dataset, (d)-(f) are our fusion results on the CT-MRI dataset.

4.4. Analysis of computational complexity

As shown in Table 4, a complexity evaluation is introduced to evaluate the efficiency of our method from two aspects, *i.e.*, training parameters and runtime. It is worthy pointing that though our method does not perform the best in terms of the model complexity and inference time due to subtle design of various modules, the proposed CTFusion and the best SOTA method are still level pegging. This indicates the efficiency of our CTFusion, which can serve practical vision tasks well with better visual performance.

Table 4. Computational efficiency comparison of five SOTA methods, the value is tested on GPU.

Methods	IFCNN[29]	PMGI[30]	CrossFuse[31]	RFN-Nest[32]	FusionGAN[21]	Ours
SIZE(M)	0.084	0.042	1.161	30.097	0.926	2.119
TIME(s)	0.013	0.052	1.076	0.358	1.179	0.019

5. Ablation studies

In the ablation study, we demonstrate the effectiveness of the self-supervised mask reconstruction task, CNN-Transformer-based encoder and the proposed S-CPFF. The experimental results are shown in Figure 9 and Table 5.

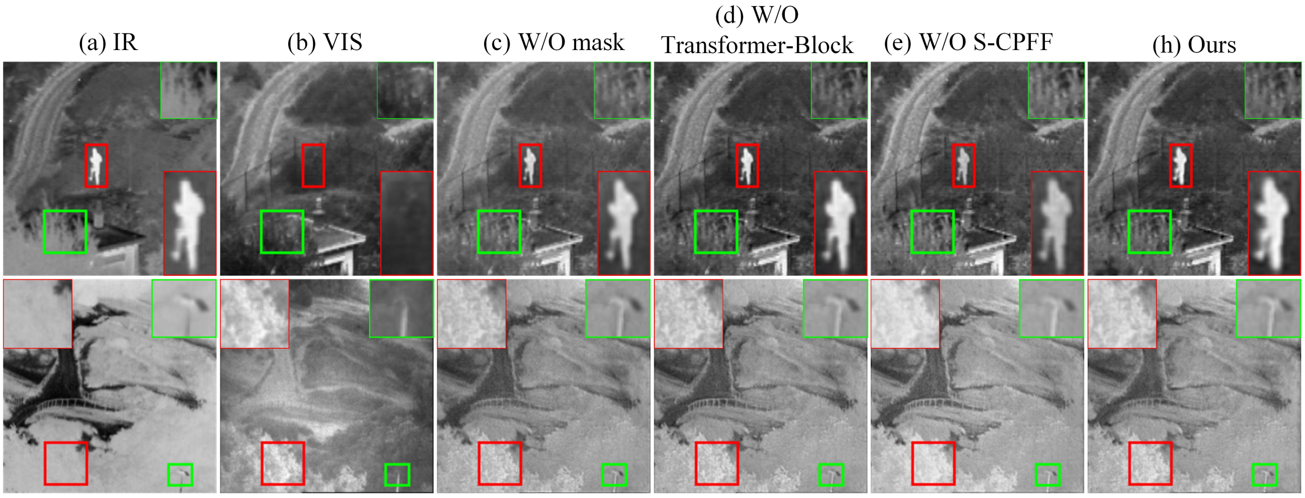


Figure 9. Vision quality comparison of the ablation study on proposed modules. From left to right, infrared image, visible image, and the results of W/O self-supervised mask reconstruction task, Transformer-Block, S-CPFF, and our CTFusion.

Table 5. Quantitative evaluation results of ablation study on 30 pairs of infrared and visible images from TNO dataset.

Configuration	CC	CE	Q^{CV}	SCD	SSIM
w/o mask	0.653	1.571	485.9	*1.427*	0.626
w/o Transformer-Block	*0.669*	1.534	499.5	1.413	0.651
w/o S-CPFF	0.655	*1.522*	*479.8*	1.388	*0.665*
Ours	**0.675**	**1.499**	**433.6**	**1.476**	**0.701**

Firstly, we remove the mask reconstruction pretext task, simply training a complete encoder-decoder framework. The results show that our proposed self-supervised mask reconstruction task can improve the ability of the framework to excavating intrinsic information. To verify the effectiveness of CNN-Transformer-based encoder, we conduct an ablation study where the encoders only contain CNN-Block. From the results we can see that regardless of whether the proposed self-supervised mask reconstruction task is used, adding Transformer-Block in the encoders always improves the fusion performance. To further prove that our proposed S-CPFF is effective, we replace the fusion net with simple feature concatenation operation. Ablation study result also shows that S-CPFF highlight salient regions in source images and further promise the enhancement of texture details.

6. Discussion

Source images in this paper are all registered before fusion, which is a common data preprocessing step in IVIF task. However, in practical scenarios, although the source images can be aligned to a certain extent by carefully adjusting the installation positions of infrared and visible light sensors, it stays impossible to achieve accurate alignment directly by manual installation [34, 37]. In other words, images captured by different sensors are difficult to be strictly aligned on pixel level. In the future, we will devote to research on misaligned infrared and visible image fusion. Since the source images are misaligned and of different modalities, we need to reduce the modality discrepancy between them, so that the feature alignment can be achieved more easily. Once the features are aligned, the fusion process will not be a problem.

7. Conclusions

In this paper, we present CTFusion, a CNN-Transformer-based IVIF framework via self-supervised mask reconstruction. The CNN-Transformer-based encoder integrates the advantages of both CNN and transformer so that the network can focus on both local and global information, better understanding dependencies in images. In addition, the designed mask reconstruction task is naturally adaptive to intrinsic information excavation requirement in IVIF. Extensive experiments on three infrared-visible image datasets demonstrate the effectiveness of the proposed method.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article. All authors reviewed the manuscript.

Acknowledgments

This work was supported by 2023 Opening Research Fund of Yunnan Key Laboratory of Digital Communications (YNJTKFB-20230686, YNKLDC-KFKT-202301).

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.

References

1. Y. Liu, X. Chen, Z. Wang, Z. Wang, R. Ward and X. Wang, Deep learning for pixel-level image fusion: recent advances and future prospects, *Information Fusion*, **42** (2018), 158–173. https://doi.org/10.1016/j.inffus.2017.10.007
2. H. Zhang, H. Xu, X. Tian, J. Jiang and J. Ma, Image fusion meets deep learning: a survey and perspective, *Information Fusion*, **76** (2021), 323–336. https://doi.org/10.1016/j.inffus.2021.06.008
3. J. Ma, Y. Ma and C. Li, Infrared and visible image fusion methods and applications: a survey, *Information Fusion*, **45** (2018), 153–178. https://doi.org/10.1016/j.inffus.2017.02.004
4. C. Yang, J. Zhang, X. Wang, and X. Liu, A novel similarity based quality metric, *Information Fusion*, **9** (2008), 156–160. https://doi.org/10.1016/j.inffus.2006.09.001
5. L. Qu, S. Liu, and M. Wang, Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning, *Proceedings of the AAAI conference on artificial intelligence*, (2022), 2126–2134. https://doi.org/10.48550/arXiv.2112.01030

6. X. Zhang, P. Ye and G. Xiao, VIFB: a visible and infrared image fusion benchmark, *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2020), 468–478. https://doi.org/10.1109/CVPRW50498.2020.00060

7. L. Sun, Y. Li, M. Zheng, Z. Zhong, and Y. Zhang, Mcnet: Multiscale visible image and infrared image fusion network, *Signal Processing*, **208** (2023), 108996. https://doi.org/10.1016/j.sigpro.2023.108996

8. L. Chipman, T. Orr, and L. Graham, Wavelets and image fusion, *Proceedings of the International Conference on Image Processing*, (1995), 248–251.

9. A. V. Vanmali, and V. M. Gadre, Visible and nir image fusion using weight-map-guided Laplacian–Gaussian pyramid for improving scene visibility, *Sādhanā*, **42** (2017), 1063–1082.

10. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

11. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You only look once: unified, real-time object detection, *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91

12. O. Ronneberger, P. Fischer and T. Brox, U-net: convolutional networks for biomedical image segmentation, *Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

13. K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, Masked Autoencoders Are Scalable Vision Learners, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

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

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

16. Y. Niu, S. Xu, L. Wu, and W. Hu, Airborne infrared and visible image fusion for target perception based on target region segmentation and discrete wavelet transform, *Mathematical Problems in Engineering*, **2012** (2012), 732–748. https://doi.org/10.1155/2012/275138

17. D. Bulanon, T. Burks, and V. Alchanatis, Image fusion of visible and thermal images for fruit detection, *Biosyst. Eng*, **103** (2009), 12–22.

18. M. Choi, R. Kim, M. Nam, and H. Kim, Fusion of multispectral and panchromatic satellite images using the curvelet transform, *IEEE Geosci. Remote Sensing Lett*, **2** (2005), 136–140. https://doi.org/10.1109/LGRS.2005.845313

19. Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang Infrared and visible image fusion with convolutional neural networks, *Multiresolution and Information Processing*, **16** (2018), 1850018. https://doi.org/10.1142/S0219691318500182

20. H. Li, X. Wu, DenseFuse: A fusion approach to infrared and visible images, *IEEE Transactions on Image Processing*, **28** (2018), 2614–2623. https://doi.org/10.1109/TIP.2018.2887342

21. J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, Fusiongan: A generative adversarial network for infrared and visible image fusion, *Information Fusion*, **48** (2019), 11–26. https://doi.org/10.1016/j.inffus.2018.09.004

22. H. Zhang, and J. Ma , SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, *International Journal of Computer Vision*, **129** (2021), 2761–2785. https://doi.org/10.1007/s11263-021-01501-8

23. H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, U2fusion: A unified unsupervised image fusion network, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44** (2020), 502–518. https://doi.org/10.1109/TPAMI.2020.3012548

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

25. J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 5802–5811. https://doi.org/10.1109/CVPR52688.2022.00571

26. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, Residual dense network for image restoration, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43** (2020), 2480–2495. https://doi.org/10.1109/TPAMI.2020.2968521

27. R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, VIF-Net: an unsupervised framework for infrared and visible image fusion, *IEEE Transactions on Computational Imaging*, **6** (2020), 640–651. https://doi.org/10.1109/TCI.2020.2965304

28. J. Liu, Y. Wu, Z. Huang, R. Liu, and X. Fan, Smoa: Searching a modality-oriented architecture for infrared and visible image fusion, *IEEE Signal Processing Letters*, **28** (2021), 1818–1822. https://doi.org/10.1109/LSP.2021.3109818

29. Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, IFCNN: A general image fusion framework based on convolutional neural network, *Information Fusion*, **54** (2020), 99–118. https://doi.org/10.1016/j.inffus.2019.07.011

30. H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 12797–12804. https://doi.org/10.1609/aaai.v34i07.6975

31. H. Li, and X. Wu, CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach, *Information Fusion*, **103** (2024), 1566–2535. https://doi.org/10.1016/j.inffus.2023.102147

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

33. D. P. Kingma, and J. Ba, Adam: A method for stochastic optimization, *arXiv: Learning*, **73** (2014).

34. H. Li, J. Liu, Y. Zhang, and Y. Liu, A Deep Learning Framework for Infrared and Visible Image Fusion Without Strict Registration, *International Journal of Computer Vision*, (2023). https://doi.org/10.1007/s11263-023-01948-x

35. A. Toet, Tno image fusion dataset, https://doi.org/10.6084/m9.figshare.1008029.v2

36. H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion, *IEEE Transactions on Image Processing*, **30** (2021), 4070–4083. https://doi.org/10.1109/TIP.2021.3069339

37. H. Li, J. Zhao, J. Li, Z. Yu and G. Liu Feature Dynamic Alignment and Refinement for Infrared-Visible Image Fusion: Translation Robust Fusion, *Information Fusion*, **95** (2023), 26–41. https://doi.org/10.1016/j.inffus.2023.02.011

38. H. Li, X. Qi, and W. Xie Fast infrared and visible image fusion with structural decomposition, *Knowledge-Based Systems*, **204** (2020), 106182. https://doi.org/10.1016/j.knosys.2020.106182

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

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

41. H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, and D. Tao Discriminative Dictionary Learning-Based Multiple Component Decomposition for Detail-Preserving Noisy Image Fusion, *IEEE Transactions on Instrumentation and Measurement*, **69** (2020), 1082–1102. https://doi.org/10.1109/TIM.2019.2912239

42. W. Xiao, Y. Zhang, H. Wang, F. Li, and H. Jin Heterogeneous Knowledge Distillation for Simultaneous Infrared-Visible Image Fusion and Super-Resolution, *IEEE Transactions on Instrumentation and Measurement*, **71** (2022), 5004015. https://doi.org/10.1109/TIM.2022.3149101

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

44. Y. Zhang, Y. Wang, H. Li, and S. Li Cross-Compatible Embedding and Semantic Consistent Feature Construction for Sketch Re-identification, *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, (2022), 3347–3355. https://doi.org/10.1145/3503161.3548224

45. H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi Triple Adversarial Learning and Multi-view Imaginative Reasoning for Unsupervised Domain Adaptation Person Re-identification, *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2022), 2814-2830. https://doi.org/10.1109/TCSVT.2021.3099943

46. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, G. Aidan, Łu. Kaiser, and I. Polosukhin Attention is All you Need, *Advances in Neural Information Processing Systems*, 30(2017).

47. H. Chen, Y. Wang, and T. Guo Pre-Trained Image Processing Transformer, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 12299–12310.

48. X. Zhu, W. Su, and L. Lu Deformable Detr: Deformable Transformers for End-to-End Object Detection, *https://arxiv.org/abs/2010.04159*.

49. S. Zheng, J. Lu, and H. Zhao Rethinking Semantic Segmentation from A Sequence-to-Sequence Perspective with Transformers, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 6881–6890.

50. A. Dosovitskiy, L. Beyer, A. Kolesnikov An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *Proceedings of the International Conference on Learning Representations (ICLR)*.

51. K. Han, A. Xiao, and E. Wu Transformer in Transformer, *Advances in Neural Information Processing Systems*, 34(2021), 15908–15919.

52. C. Chen, R. Panda, and Q. Fan Regionvit: Regional-to-Local Attention for Vision Transformers, *https://arxiv.org/abs/2106.02689*.

53. V. Vibashan, J. Valanarasu, and P. Oza Image Fusion Transformer, *https://arxiv.org/abs/2107.09011*.

54. X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou LLVIP: A Visible-infrared Paired Dataset for Low-light Vision, *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3489-3497. https://doi.org/10.1109/ICCVW54120.2021.00389

55. *Harvard medical website. http://www.med.harvard.edu/AANLIB/home.html.7*.

56. D. Manjusha, and B. Udhav Image fusion and image quality assessment of fused images, *International Journal of Image Processing (IJIP)*, 4(2010), 484.

57. H. Chen, and P. Varshney A human perception inspired quality metric for image fusion based on regional information, *Information fusion*, 8(2007), 193–207.

58. V. Aslantas, and E. Bendes A new image quality metric for image fusion: The sum of the correlations of differences, *Aeu-international Journal of electronics and communications*, 69(2015), 1890–1896.

59. Z. Wang, and A. Bovik A universal image quality index, *IEEE Signal Processing Letters*, 9(2002), 81–84.

AIMS Press