Information Fusion 103 (2024) 102147

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach

Hui Li *, Xiao-Jun Wu

International Joint Laboratory on Artificial Intelligence of Jiangsu Province, School of Artificial Intelligence and Computer Science, Jiangnan University, 214122, Wuxi, China

ARTICLE INFO

Keywords:
Image fusion
Transformer
Cross attention
Infrared image
Visible image

ABSTRACT

Multimodal visual information fusion aims to integrate the multi-sensor data into a single image which contains more complementary information and less redundant features. However the complementary information is hard to extract, especially for infrared and visible images which contain big similarity gap between these two modalities. The common cross attention modules only consider the correlation, on the contrary, image fusion tasks need focus on complementarity (uncorrelation). Hence, in this paper, a novel cross attention mechanism (CAM) is proposed to enhance the complementary information. Furthermore, a two-stage training strategy based fusion scheme is presented to generate the fused images. For the first stage, two auto-encoder networks with same architecture are trained for each modality. Then, with the fixed encoders, the CAM and a decoder are trained in the second stage. With the trained CAM, features extracted from two modalities are integrated into one fused feature in which the complementary information is enhanced and the redundant features are reduced. Finally, the fused image can be generated by the trained decoder. The experimental results illustrate that our proposed fusion method obtains the SOTA fusion performance compared with the existing fusion networks. The codes of our fusion method will be available soon.

1. Introduction

The aim of image fusion is to improve the visual quality of images and provide more accurate and reliable information for various applications [1,2]. Multimodal image fusion plays a very important role in computer vision field and industrial area [3,4], which involves combining information from different imaging modalities such as visible, infrared etc. Specially, with the development of vision sensors, how to utilize the benefits of these multimodality data to sever the real world scenario becomes a crucial problem. To this end, a lot of researchers work on the visual information fusion task and many milestone achievements have been made, such as multi-scale transformer [5,6], sparse representation [7,8], pre-trained deep learning models [9,10], Bayesian based fusion model [11]. Multimodal image fusion also has numerous applications in other fields such as medical diagnosis [12,13], surveillance [14,15], remote sensing [16,17], and robotics [18,19] etc.

Deep learning, as a widely popular technology, is no exception in multimodal image fusion field [20–24]. Thanks to the multimodal datasets [25,26], the deep learning based fusion methods have emerged as a promising approach due to their ability to learn complex feature representations from source images and effectively fuse multiple images. These methods can be broadly classified into two categories: multi-stage fusion and end-to-end fusion.

Multi-stage fusion methods [20,27–29] involve separate processing of the input images before combining them, whereas end-to-end fusion methods [24,30]combine the input images directly. Both types of methods have their strengths and disadvantages. Multi-stage fusion methods tend to be more flexible and adaptable, allowing for the incorporation of a wide range of pre-processing techniques. However, they are often computationally intensive and can suffer from information loss during the processing stages.

Multi-stage image fusion methods divide the fusion process into several stages [20,28]. Each stage handles a particular aspect of the fusion process, such as feature extraction, feature fusion and image generation. In this kind of approaches, deep learning models can be injected in each stage, such as convolutional neural networks (CNNs) for feature extraction [7,9,31], a specific light network is trained to replace the handcrafted fusion strategy [28]. These methods tend to be more flexible and adaptable, allowing for the incorporation of a wide range of pre-processing techniques. However, it requires careful selection of the models, fusion strategies and training strategies to ensure optimal results.

* Corresponding author.
E-mail address: lihui.cv@jiangnan.edu.cn (H. Li).

https://doi.org/10.1016/j.inffus.2023.102147
Received 5 September 2023; Received in revised form 11 November 2023; Accepted 13 November 2023
Available online 15 November 2023

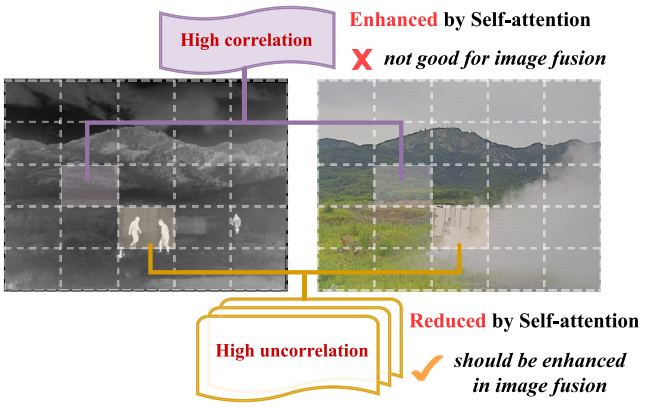

Fig. 1. The correlation and uncorrelation for self-attention in image fusion task. For multimodal images, the self-attention may not be suitable for inter-modality processing. In image fusion task, the redundant information will be enhanced and the complementary features are reduced, which is more obvious when the source images are all in gray-scale.

End-to-end fusion methods [30,32], on the other hand, are more efficient and can preserve more information without the manual operations. These methods involve designing a single deep learning model that takes multiple input images and directly generates fused image. The fusion networks are utilized to extract features and learn how to fuse them in a single step [33,34]. One of the advantages of these approaches is that the optimal fusion strategy can be learned from the input images without requiring any prior knowledge of the fusion process. However, it can be more challenging to design an effective end-to-end fusion model, and it requires a carefully designed loss function to achieve optimal performance [35]. Unfortunately, these disadvantages are also the crucial problems in most computer vision tasks.

To address the above disadvantages, transformer architecture, as a key technology, has been introduced into multimodal image fusion task [22,29,36,37]. Transformer-based methods have shown promising performance in various tasks, including natural language processing [38] and image classification [39,40]. These approaches use self-attention mechanism to capture global dependencies and facilitate efficient feature representation learning. Current transformer-based fusion methods also follow this scheme. However, these methods primarily concentrate on the self-attention mechanism and ignore the interplay between various modalities [29,34,41,42]. However, the complementary information between different modalities is the key for the multimodal fusion task, thus cross-attention should be paid more attention.

Correlation among samples is a fundamental aspect of computer vision fields, as it reflects the significant features for various tasks [43–45]. However, in image fusion fields, particularly multimodal image fusion, complementary (uncorrelation) information is crucial [3,4]. Therefore, the uncorrelation should be paid more attention in image fusion tasks. Current transformer-based fusion methods [37,42,46] focus solely on the self-attention mechanism, which is the primary component of transformer. As shown in Fig. 1, while this mechanism can improve the correlation between inputs, it may also reduce the complementary information.

In recent studies, while the self-attention mechanism has demonstrated the capability to enhance complementary information through well-designed loss functions and feature fusion modules, it is important to note that mishandling feature correlation can result in a degradation of fusion performance in specific scenarios.

To overcome the limitations of current transformer-based fusion techniques, in this paper, a novel cross-attention mechanism (CAM) based fusion method is proposed, which employs self-attention to enhance the intra-features of each modality while utilizes cross-attention based architecture to enhance the inter-features (complementary information) between different modalities. By injecting the CAM into the transformer architecture, the proposed method offers a powerful fusion strategy for multimodal images that effectively enhances the uncorrelation between inputs. The main contributions of this paper are summarized as follows,

1. A cross-attention mechanism is introduced to enhance multimodal features in this paper. The proposed mechanism optimizes the fusion process by effectively augmenting complementary features, resulting in outcomes that are both more accurate and comprehensive.

2. A novel hybrid fusion network is presented in this research, amalgamating the strengths of convolutional layers with attention mechanisms (both self and cross) for the multimodal image fusion task. This methodology facilitates the extraction of deep features from source images, maintaining detail information, and enhancing complementary information.

3. In comparison to state-of-the-art fusion methods, the experimental results demonstrate that the method proposed in this paper presents a promising alternative to current fusion techniques. It furnishes a more robust and efficient solution for the multimodal image fusion task.

The rest of our paper is structured as follows. In Section 2, we briefly review the related work on deep learning-based fusion. The proposed fusion framework is described in detail in Section 3. The experimental results are presented in Section 4. Finally, we draw the paper to conclusion in Section 5.

2. Related works

In this section, two key techniques based methods are briefly introduced, including: transformer based fusion methods and cross-attention based methods.

2.1. Transformer based fusion methods

Transformer is a deep learning architecture originally developed for natural language processing tasks [38,47], it has also been successfully applied to computer vision fields [39,41,46]. The key innovation is the self-attention mechanism, which allows the model to weigh the importance of different input elements when making predictions.

In computer vision, the transformer can be used in various ways. One common approach is to use it as an alternative to CNNs for feature extraction, it also appears in image fusion task [48,49]. Instead of using a fixed kernel to extract local features, the transformer computes attention weights between all pairs of input elements, allowing it to capture global dependencies and relationships among the features. This approach has also been shown to be effective for image fusion tasks [50,51].

Although these transformer-based fusion methods obtain good fusion performance, the drawbacks are still observed: (1) The transformer architecture is only utilized in feature extraction stage or image reconstruction stage, the relations between different modalities are not considered [29,41]; (2) Even with the self-attention mechanism in feature fusion stage, these methods still do not address the crucial problem which is that the self-attention mechanism may reduce the complementary information [22,52].

2.2. Cross-attention based fusion methods

Cross-attention is a technique used in computer vision tasks that primarily focuses on the interaction of information between different modalities [53–55]. It is often utilized in multimodal tasks where information from different modalities needs to be integrated to solve a specific problem, such as image fusion [56–58], and image registration based fusion method [59,60].

In transformer architecture, cross-attention is also a key concept, which has been shown to be effective for integrating information

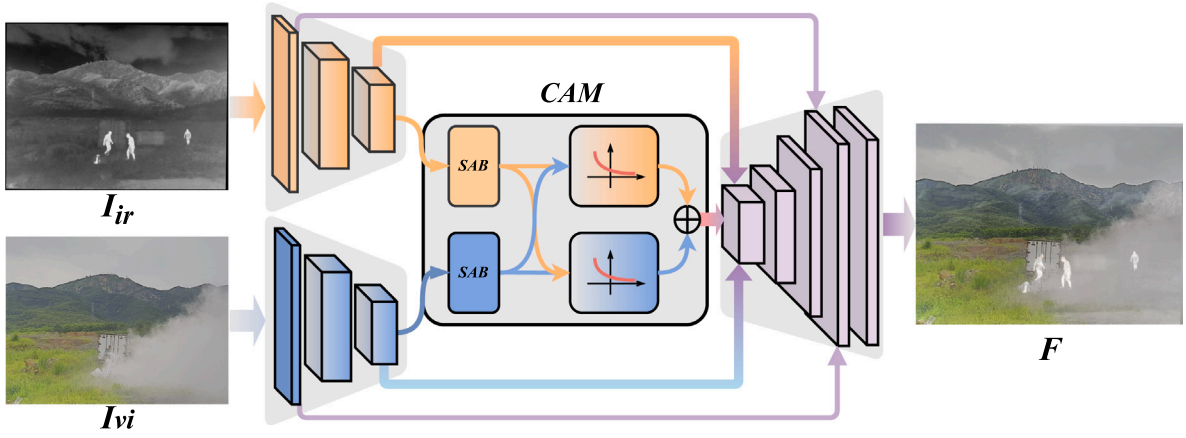

Fig. 2. The framework of CrossFuse. Two "Encoder" contain same architecture but different parameters. The cross-attention mechanism (CAM) is utilized to fuse the multimodal features. "SAB" indicates the self-attention block. The fused image can be obtained by "Decoder" with the long connection from encoders.

from multiple modalities [54,61]. In transformer, cross-attention is computed between the features of the encoder and the decoder, where the encoder produces one modality features and the decoder produces another modality features. There are also some fusion methods combine the cross-attention and transformer to obtain better performance [22, 52]. However, these cross-attention only focuses on the correlation and ignores the complementary information.

Although the cross-attention mechanism has received widespread attention in image fusion task, the relationship between attention mechanism and the fundamental issue of fusion task has not been fully explored. Thus, how to design an appropriate cross-attention mechanism which preserves the complementary information is crucial.

3. The proposed method

The proposed CAM based fusion network focuses on the fundamental problem of image fusion task, in which the cross-attention mechanism in image fusion task should enhance the complementary (uncorrelation) information and reduce the redundant (correlation) features. In this section, the architecture and the loss function are introduced in detail.

3.1. The architecture of the fusion network

The architecture is shown in Fig. 2. I_{ir} and I_{vi} indicate the infrared image and visible image, respectively. Two encoders are utilized to extract multimodal features from source images. The CAM based transformer architecture is introduced to fuse the multimodal features. Finally, the fused image is generated by decoder. There are two skip connections between encoder and decoder, which are utilized to preserve more deep features and shallow features from source images.

3.1.1. Encoder architecture

Considering the gap between two modalities (infrared and visible), it is natural to extract the features with different parameters. Thus, in our framework, two encoders with the same architecture but different parameters are utilized. The architecture of encoder is shown in Fig. 3.

The first convolutional layer, "Conv", is utilized to extract the shallow features from source images, which contains rich texture information. Following the pooling operation, MaxPooling, and the multi-scale features are exploited and the features will preserve more useful information with DenseBlock. With the deeper of encoder, the extracted deep features will focus on salient contents.

Furthermore, to enhance the detail information and salient features, two skip connections (Conv and the last DenseBlock follow the MaxPooling) are applied into encoder and decoder.

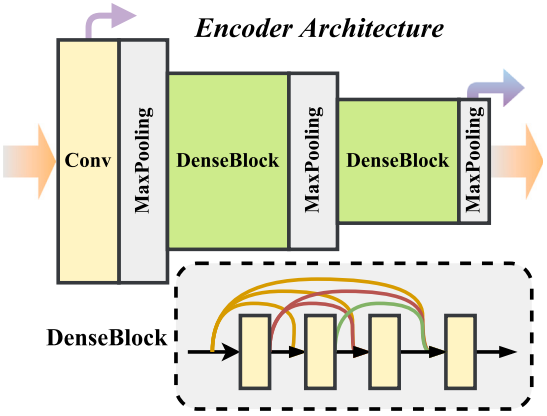

Fig. 3. The encoder architecture which contains three blocks: "Conv", "MaxPooling" and "DenseBlock". "Conv" indicates one convolutional layer, "DenseBlock" includes four convolutional layers with dense connection.

3.1.2. Cross-attention mechanism (CAM)

The proposed cross-attention mechanism (CAM) is introduced in this section, which is the most important part in our method. The architecture of CAM is shown in Fig. 4.

Two branches with different parameters are utilized to extract features from two modalities. Each modality features are fed to self-attention (SA) block first to enhance the intra-features, which fits the insight of SA. To further enhance the intra-features, the shift operation is also introduced into CAM, in which the positions of features are moved horizontally and vertically. Then, another SA block is used to enhance the shifted features which will contain more global information. Before the cross-attention, the "unshift" is utilized to restore the positions. Thus, there are twice as many SA as CA.

After obtaining the intra-enhanced features, the proposed cross-attention block is introduced. The formulas of SA are given as follows,

$$\begin{aligned} &[Q_c, K_c, V_c] = x^c U_{qkv}, \\ &x^c_{sa} = x^c_{sa} + norm(softmax(\frac{Q_c K_c^T}{\sqrt{d}})V_c), \\ &x^c_{sa} = x^c_{sa} + MLP(norm(x^c_{sa})), \\ &s.t. \quad U_{qkv} \in \mathbb{R}^{d\times 3d}, c \in \{ir, vi\} \end{aligned} \tag{1}$$

where x^c means the input of SA, Q_c, K_c and V_c indicate the different representation of input, d is the dimension of the input vector. U_{qkv} is a transformation matrix which can be learned by a fully connection

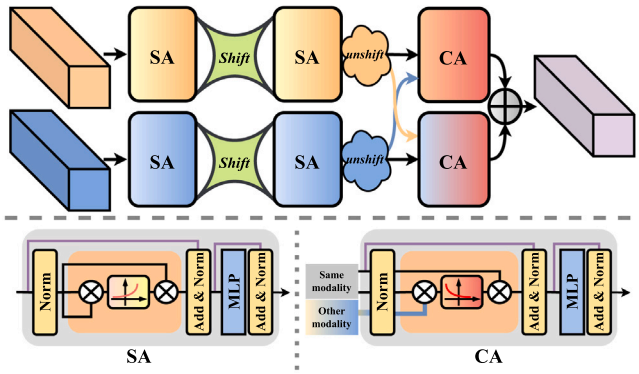

Fig. 4. The cross-attention mechanism architecture. "SA" follows the standard transformer architecture which contains one self-attention block. The "Shift" and "unshift" mean the block shift and shift back operation. "CA" indicates the novel cross-attention mechanism which focuses on the uncorrelation information.

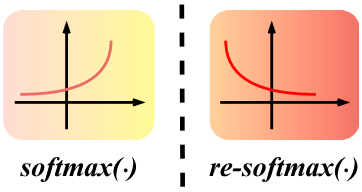

Fig. 5. The activation function curves of $softmax(\cdot)$ and $re\text{-}softmax(\cdot)$.

layer, $norm(\cdot)$ indicates the linear norm operation, $MLP(\cdot)$ means the multilayer perceptron.

The formulas of CA are given as follows,

$$\begin{aligned}
&[Q_{\hat{c}}, K_c, V_c] = [x^{\hat{c}}, x^c, x^c]\mathcal{U}_{qkv},\\
&x^c_{ca} = x^c_{ca} + norm(re\text{-}softmax(\frac{Q_{\hat{c}}K_c^T}{\sqrt{d}})V_c),\\
&x^c_{ca} = x^c_{ca} + MLP(norm(x^c_{ca})),\\
&s.t. \quad \mathcal{U}_{qkv} \in \mathbb{R}^{3d\times 3d}, c \in \{ir, vi\}, \hat{c} \neq c
\end{aligned} \tag{2}$$

where the $\hat{c}$ and c indicate the different modality.

The main difference between SA and CA is that the activation function after the matrix multiplication. For different modalities, the complementary (uncorrelation) information rather than the redundant (correlation) features should be enhanced. Thus, a new activation function, reversed softmax ($re\text{-}softmax$), is embed into our cross-attention mechanism, the formulation is given as follows,

$$re\text{-}softmax(X) = softmax(-X) \tag{3}$$

The activation function curves of $softmax(\cdot)$ and $re\text{-}softmax(\cdot)$ are shown in Fig. 5 which shows the trend of different activation functions. With the $re\text{-}softmax(\cdot)$, our CA block can focus on the uncorrelation information between different modalities.

3.1.3. Decoder architecture

After CAM, a decoder network is introduced into our framework to obtain the final fused image. In this decoder, several convolutional layers and up-sampling operations are included. The architecture is shown in Fig. 6.

To better enhance the salient features and preserve more detail information from source image features, two skip connections between encoder and decoder are introduced into our network, where the deep feature connection for salient features and the shallow connection for detail information. In addition, the feature intensity aware strategy is applied into decoder for multi-level feature fusion, the formula is

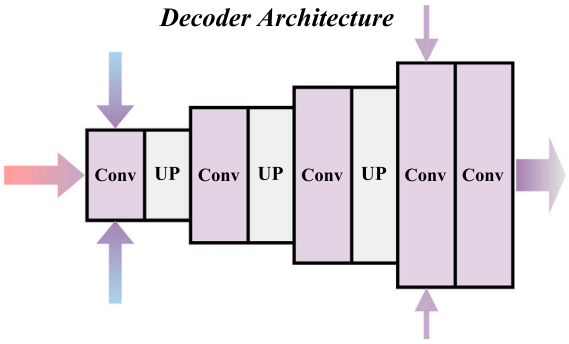

Fig. 6. The decoder architecture.

defined as follows,

$$\begin{aligned}
&\Phi^m_{df} = \Phi^m_c + w^m_{ir}\Phi^m_{ir} + w^m_{vi}\Phi^m_{vi}, \quad m \in \{deep, shallow\}\\
&s.t. \quad w^m_{ir/vi}(\cdot) = \frac{\nabla^m \Phi^m_{ir/vi}(\cdot)}{\sum_{\hat{i}\in\{ir,vi\}} \nabla^m \Phi^m_{\hat{i}}(\cdot)}
\end{aligned} \tag{4}$$

where $(\cdot)$ means the position in deep features, Φ^m_c indicates the features extracted by CAM, Φ^m_{ir} and Φ^m_{vi} denote the features from source images (infrared and visible). ∇^m denotes the detail and base information extractor for shallow features and deep features, respectively. The formulas of ∇^m are given as follows,

$$\begin{aligned}
&\nabla^{deep}\Phi = k_\nabla \otimes \Phi\\
&\nabla^{shallow}\Phi = \sqrt{(1 - k_\nabla \otimes \Phi)^2}
\end{aligned} \tag{5}$$

$$k_\nabla = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1 \end{bmatrix} \tag{6}$$

where $\otimes\, k_\nabla$ is the kernel of convolutional operation to extract the base information. Given Eqs. (5) and (6), ∇^{deep} focuses on the salient features and $\nabla^{shallow}$ enhances the detail information.

3.2. Training phase with two-stage strategy

To train our fusion framework, a two-stage training strategy [28] is applied. Firstly, an auto-encoder network is constructed for each modality (infrared and visible), which is utilized to reconstruct the inputs. Then, with the trained encoders for each modality, the proposed CAM and decoder are trained with the multimodal data and the proposed loss function.

3.2.1. First stage for encoders

In first stage, the encoders are trained to extract rich features which are benefit for generating the fused image, the framework is shown in Fig. 7. Since there are feature gap between infrared and visible, it is reasonable to train different parameters for each modality.

As shown in Fig. 7, these two auto-encoders have same network structure but different parameters. Two skip connections are utilized to preserve the shallow features (detail) and deep features (salient).

Furthermore, to train the auto-encoder network, pixel-level loss ($\|\cdot\|_F^2$) and structural similarity loss ($SSIM$) are introduced. The loss function for auto-encoders is given as follows,

$$L^c_{auto} = \|I_c - I^r_c\|_F^2 + w_s SSIM(I_c, I^r_c), \quad c \in \{ir, vi\} \tag{7}$$

where I^r_c indicates the reconstructed image along with the certain modality(infrared or visible), w_s denotes the trade-off parameter which is set to $1e4$.

Finally, we only use the trained encoders to extract deep features from corresponding modality.

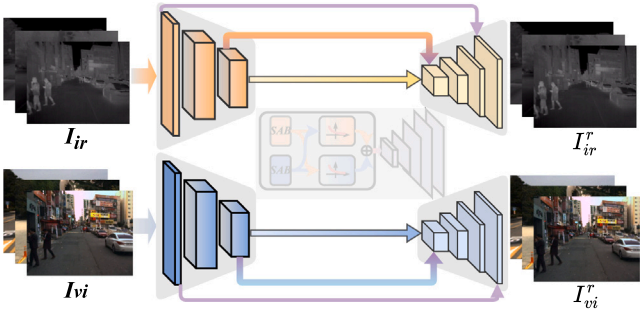

Fig. 7. The first training strategy. Two auto-encoders are trained to reconstruct the inputs (infrared images and visible images).

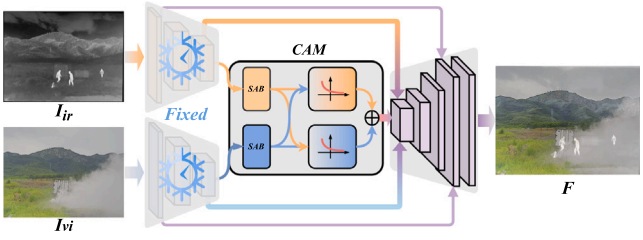

Fig. 8. The second training strategy.

3.2.2. Second stage for CAM and decoder

In second training stage, with the fixed encoders, the proposed CAM and the decoder are trained. As shown in Fig. 8, between the fixed encoders and the decoder, two skip connections are also applied into the final stage.

Since the fused image should contain more complementary features and reduce redundant information from different modalities, a novel attention-based loss function is proposed to train our network. The formula of our loss function is given as follows,

$$L_{cam} = L_{int} + w_g L_{gra} \tag{8}$$

where w_g indicates the trade-off parameter between two terms, which is set to 10. L_{int} and L_{gra} denote the intensity loss and the gradient loss.

The intensity loss function: The pixel intensity indicates the main part of fused image, such as illumination, contour etc. Usually, these information do not always appear in single modality. Thus, the intensity mask is introduced into this loss function, which is given as follow,

$$L_{int} = \|F - (M_{ir} I_{ir} + M_{vi} I_{vi})\|_F^2 \tag{9}$$

where F indicates the fused image, M_{ir} and M_{vi} denote the intensity masks for different modality.

The masks are calculated as follows,

$$M_{ir} = \begin{cases} 1, & loc_{ir} \geq loc_{vi} \\ 0, & otherwise \end{cases}, M_{vi} = 1 - M_{ir} \tag{10}$$

where loc_{ir} and loc_{vi} indicate the mean value of local patch from source images. These values are calculated as follows,

$$loc_c = \frac{avg_c}{\sum_{i\in\{ir,vi\}} avg_i}, \quad avg_c = \nabla_a I_c \tag{11}$$

where avg_c indicates the single modality ($c \in \{ir, vi\}$) values obtained by mean filter ∇_a, in which the kernel size is 11×11.

The gradient loss function: Since the intensity loss function only focuses on the illumination and contour information, the gradient loss function is utilized to ensure that the detail information can be preserved. The formula of gradient loss function is given as follows,

$$\begin{aligned} &L_{gra} = \|F - max(Clip(\nabla_g I_{ir}), Clip(\nabla_g I_{vi}))\|_F^2 \\ &s.t. \quad Clip(\cdot) = max(\cdot, 0) \end{aligned} \tag{12}$$

Fig. 9. The examples of two datasets: TNO and VOT-RGBT.

where ∇_g denotes the mean filter with a small kernel size 3×3. The mean filter with small kernel size can extract higher robustness features and more detail information.

4. Experimental validation

In this section, the comparison experiments are conducted to evaluate the fusion performance of the proposed fusion method. After introducing the experimental settings, several ablation studies are performed to investigate the effect of different elements of the proposed fusion network. Several performance metrics are utilized to evaluate the fusion performance objectively.

Our network is implemented on the NVIDIA GPU (GTX 3090Ti) using PyTorch as a programming environment.

4.1. Experimental settings

In training phase, for the first stage (two auto-encoders), 40000 pairs of infrared and visible images are randomly chosen from the KAIST dataset. The epoch and the batch size are set to 4 and 2, respectively. For the second stage (CAM and decoder), 20000 pairs of infrared and visible images are chosen, the epoch and the batch size are set to 8 and 8, respectively. The initial learning rate is set to 0.01 and decreased by one tenth every 2 epochs. All these images are converted to gray scale and resized to 256×256.

The test images are selected from TNO [62] and VOT-RGBT [63], comprising 21 and 40 pairs of infrared and visible images, respectively. The TNO dataset encompasses more intricate scenarios where salient objects may not always be present. In contrast, the VOT-RGBT dataset primarily concentrates on street scenarios with smaller salient targets. The examples of these two datasets are shown in Fig. 9.

To evaluate the fusion performance of our proposed network, eight state-of-the-art fusion methods are chosen: a generative adversarial network (GAN) based fusion method (FusionGAN) [21], a unified CNN based method (IFCNN) [32], a unified dense connection based fusion method (U2Fusion) [30], two transformer based fusion networks (YDTR [34], DATFuse [42]), a joint down-stream tasks (saliency object detection) fusion methods (IRFS) [35], a semantic based fusion methods (SemLA) [60], and a diffusion model based fusion network (DDFM) [64].

Furthermore, six image quality metrics are utilized to assess the objective evaluation, which includes: Entropy (En) [65]; Standard Deviation (SD) [66]; Mutual Information (MI) [67]; Image feature based Mutual Information (FMI_{dct}, FMI_{pixel}) [68]; the sum of correlations of differences (SCD) [69].

4.2. Ablation study

In this section, we will analyze the influence of each key part: the number of attention block, the *re-softmax* operation, the shift operation and the CAM. Furthermore, the loss function, the fusion module and the training strategy are also analyzed.

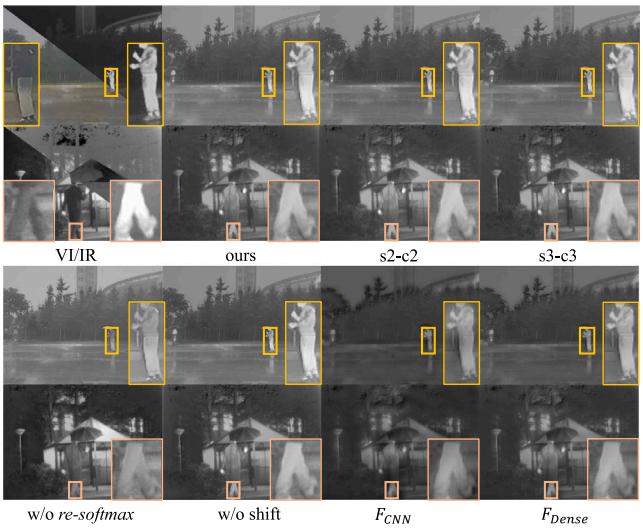

Fig. 10. The results of ablation studies with different settings. The visible image is converted to gray-scale.

Table 1
The objective results of ablation studies. "w/o cross" means the re-softmax() operation is replaced by softmax() in CA. "w/o shift" indicates the shift operation is not used between two SA module. "s1-c1", "s2-c2" and "s3-c3" denote the number of block in SA and CA. "F_{CNN}" and "F_{Dense}" indicate that the CAM is replaced by CNN(4 conv layers) and dense architecture(1 dense block and 1 conv layer).

	EN↑	SD↑	MI↑	FMI_{dct} ↑
Ours(s1-c1)	**6.8389**	**73.4712**	**13.6779**	0.3866
s2-c2	6.7282	70.0806	13.4563	0.3915
s3-c3	6.6862	66.7879	13.3725	0.3868
w/o *re-softmax*	6.8281	72.5535	13.6563	0.3943
w/o shift	6.7037	68.6275	13.4074	**0.3974**
F_{CNN}	6.7201	69.6833	13.4403	0.3900
F_{Dense}	6.7342	70.8587	13.4684	0.3919

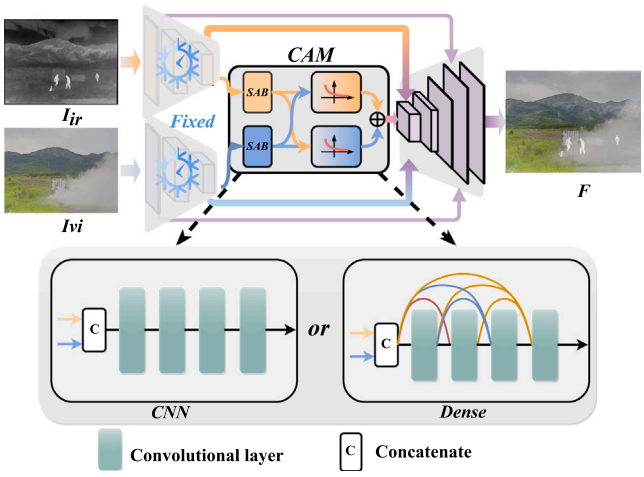

Fig. 11. The proposed network with different fusion module: CAM, CNN and Dense.

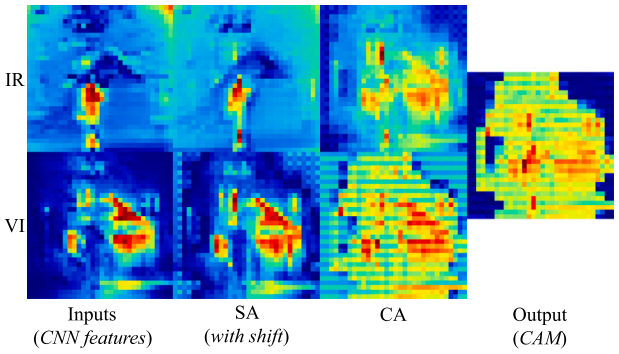

Fig. 12. The visualization of middle features obtained by CNN-based encoder, SA, CA and CAM. The size of each features is 32 × 32.

4.2.1. The number of "SA" block and "CA" block

SA and CA indicate the self-attention module and the cross module in CAM, respectively. To find the best settings of block number, the experiments of one block (s1-c1), two blocks (s2-c2) and three blocks (s3-c3) are conducted. The visualized results and the metric values are shown in Fig. 10 and Table 1, the best values are indicated in **bold**.

From Fig. 10, the result obtained by one block (ours, s1-c1) contains more detail information and less artificially generated noise around the salient object (umbrella). However, the visualized performance between these results still very close. Thus, four metrics are utilized to evaluate the performance.

In Table 1, comparing with two blocks (s2-c2) and three blocks (s3-c3), the proposed network with one SA block and one CA block (s1-c1) obtains better metric values (EN, SD, MI). Although deeper architecture has better performance in many vision tasks, it is not always correct in low-level vision task, such as image fusion. Furthermore, the proposed network is a light-weight architecture and the deep features contain less semantic features, that is why s1-c1 obtains better fusion performance.

4.2.2. The influence of *re-softmax* and shift operations

In our method, the $re\text{-}softmax(\cdot)$ operation is the key part of cross attention block, which can force network focus on the complementary (uncorrelation) information between different modalities. The shift operation is introduced to enhance the intra-features which is also applied in Swin-Transformer [40].

As shown in Fig. 10 and Table 1, in the absence of these two crucial operations, the fusion results exhibit a decrease in detail, and the intensity of salient objects is also diminished. Across the selected four metrics, our proposed scheme achieves three superior values (En, SD, MI), indicating that the $re\text{-}softmax(\cdot)$ and shift operations contribute to preserving more detailed information (En, SD) and enhancing complementary features (MI).

4.2.3. The influence of CAM

To evaluate the effectiveness of CAM, two architectures (CNN and Dense) are utilized to replace CAM in our fusion network. The architectures are shown in Fig. 11. In Fig. 10 and Table 1, F_{CNN} and F_{Dense} indicate the proposed fusion network with CNN based and dense connection based fusion module, respectively. To further analysis the influence of SA and CA in CAM, the visualization of middle features are shown in Fig. 12.

As shown in Fig. 10, comparing with F_{CNN} and F_{Dense}, the result obtained by the proposed CAM based fusion network contains more salient features and less artifacts (background), which makes the fused image more natural. Furthermore, the objective evaluation results are shown in Table 1, which also demonstrates that the CAM can improve the fusion performance combined with our feature extraction and image reconstruction network

As discussed in Section 3.1.2, the CAM is comprised of both the Self Attention (SA) module with shift and the Cross Attention (CA) module. The heatmaps of middle features are illustrated in Fig. 12. The "Inputs" are generated by the CNN-based encoder, highlighting the salient regions (as highlighted areas) following the source images. After the SA operations, as depicted in Fig. 12 (SA), not only are the salient regions retained, but the finer details (background) are also enriched within each modality. Thanks to the CA operation, the complementary regions in each branch (IR, salient parts, and VI, background) are

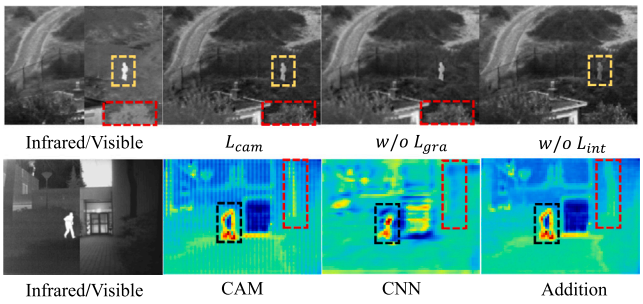

Fig. 13. Fusion results obtained by different loss function settings and the visualization of fused features obtained by different fusion strategies.

amplified through the "$re\text{-}softmax(\cdot)$". The final addition operation culminates in the acquisition of the enhanced features via our CAM. The visual results obtained by different architectures are also shown in Fig. 10 (ours, F_{CNN} and F_{Dense}).

These observations indicate that our proposed CAM effectively augments the complementary features within multimodal images, ensuring the preservation of both salient objects and detailed information.

4.2.4. Analysis for loss function and fusion module

In second training stage, the loss function (L_{cam}) contains two items: the pixel intensity part (L_{int}) and the gradient part (L_{gra}). In Fig. 13 (first row), "w/o L_{gra}" indicates only the pixel intensity part is utilized to train our network and "w/o L_{int}" means only the gradient part is used.

As shown in Fig. 13 (first row), in comparison with "L_{cam}", the fusion result obtained by "w/o L_{gra}" exhibits a reduction in detailed information, as indicated by the red box. For "w/o L_{int}", the intensity of the salient object in the fused image decreases, as highlighted in the yellow box, which is deemed unacceptable for the image fusion task. These observations underscore the effectiveness of our proposed loss function in preserving both detailed information and salient pixel intensity.

To analysis the performance of fusion module, the visual results of fused deep features are shown in Fig. 13 (second row). These heat maps are calculated by average feature maps across channel dimensions. To evaluate the fusion performance, two classical fusion modules are chosen: (1) a light-weight CNN based fusion module (CNN), and (2) additional operation based fusion module (Addition).

In Fig. 13 (second row), the feature map derived from the CNN exhibits a higher presence of redundant features. In contrast to the CNN module, the CAM proves adept at preserving more structural information from source inputs, enhancing salient objects (as highlighted in the black and red boxes). Moreover, the CAM effectively amplifies complementary regions from multi-modality inputs (as indicated by the red box), outperforming the Addition method. These observations affirm that our proposed fusion module (CAM) excels in augmenting complementary features and structural information while mitigating the presence of redundant features.

4.2.5. Analysis for training strategies

In this section, we will analyze the impact of different training strategies, namely the "two-stage" and "one-stage" strategies. "two-stage" indicates the training strategy utilized in our proposed fusion framework, "one-stage" means two encoders, CAM and decoder are trained together with the proposed loss function. The loss curve and metrics values (on TNO) are shown in Fig. 14 and Table 2.

As shown in Fig. 14, with the utilization of the proposed loss function (L_{cam}), both of these training strategies converge to a stable value. Nevertheless, under identical settings, the "two-stage" approach exhibits faster convergence and a smaller loss value compared to the "one-stage" strategy.

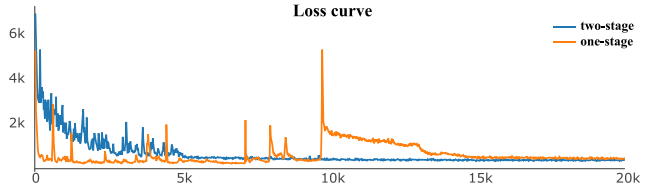

Fig. 14. Fusion results with different loss function settings and the visualization of fused features with different fusion strategies ("two-stage" and "one-stage").

Table 2
The objective results of different training strategies ("two-stage" and "one-stage") on TNO.

	EN↑	SD↑	MI↑	FMI_{dct} ↑
Two-stage	**6.8389**	**73.4712**	**13.6779**	0.3866
One-stage	6.6743	68.5877	13.3486	**0.3900**

Moreover, to ascertain the superior training strategy, we conducted comparative experiments on TNO. Table 2 reveals that the "two-stage" approach attains the three highest values, signifying that a fusion model based on the two-stage training strategy yields superior fusion performance. Based on the efficiency of training strategy and the fusion performance, in our proposed framework, we choose two-stage training strategy to train our network.

4.3. Fusion results analysis

In this section, five state-of-the-art fusion methods and six metrics are chosen to evaluate the fusion performance of our proposed fusion network. The comparison experiments are conducted on two public fusion datasets (TNO [62] and VOT-RGBT [63]). The examples "yard" and "man" from TNO, the examples "outdoors" and "two-man" from VOT-RGBT, are chosen to show the visual results.

4.3.1. Fusion results on TNO

To evaluate the fusion performance on TNO [62], 21 pairs of infrared and visible images are selected. The fusion results obtained by the proposed method and other existing fusion methods on TNO ("yard" and "man") are shown in Fig. 15.

Comparing with the GAN-based method (FusionGAN [21]), CNN-based method (IFCNN [32]) and dense connection-based method (U2Fuison [30]), the fused image obtained by our proposed method contains more detail information (Fig. 15, yellow box). Moreover, our proposed method can generate clearer fused image than the transformer-based method (YDTR [34] and DATFuse [42]) and two down-stream task based methods (IRFS [35] and SemLA [60]). For the diffusion model based fusion method (DDFM [64]), the proposed method obtains the comparable visual results on TNO.

To assess the fused image quality objectively, six metrics are selected. The metrics values are shown in Table 3, the best values are denoted in **blob** and the second-best values are denoted in *italic and red*. In Table 3, compared with other state-of-the-art fusion methods, our proposed method (CrossFuse) achieves four best values (EN, SD, MI, and FMI_{dct}) and two second-best value (FMI_{pixel} and SCD), which means the CrossFuse can preserve more complementary information from pixel-level (EN and SD) and feature-level (MI and FMI_{dct}). The above observations indicate that the CAM can maintain more complementary information from source images.

4.3.2. Fusion results on VOT-RGBT

In VOT-RGBT, 40 pairs of infrared and visible images are selected from VOT-RGBT [63] and TNO [62]. Since the visible image is in RGB space, it is converted into YCrCb color space. "Y" indicates the luminance, "Cr" and "Cb" denote the chrominance. To obtain the RGB

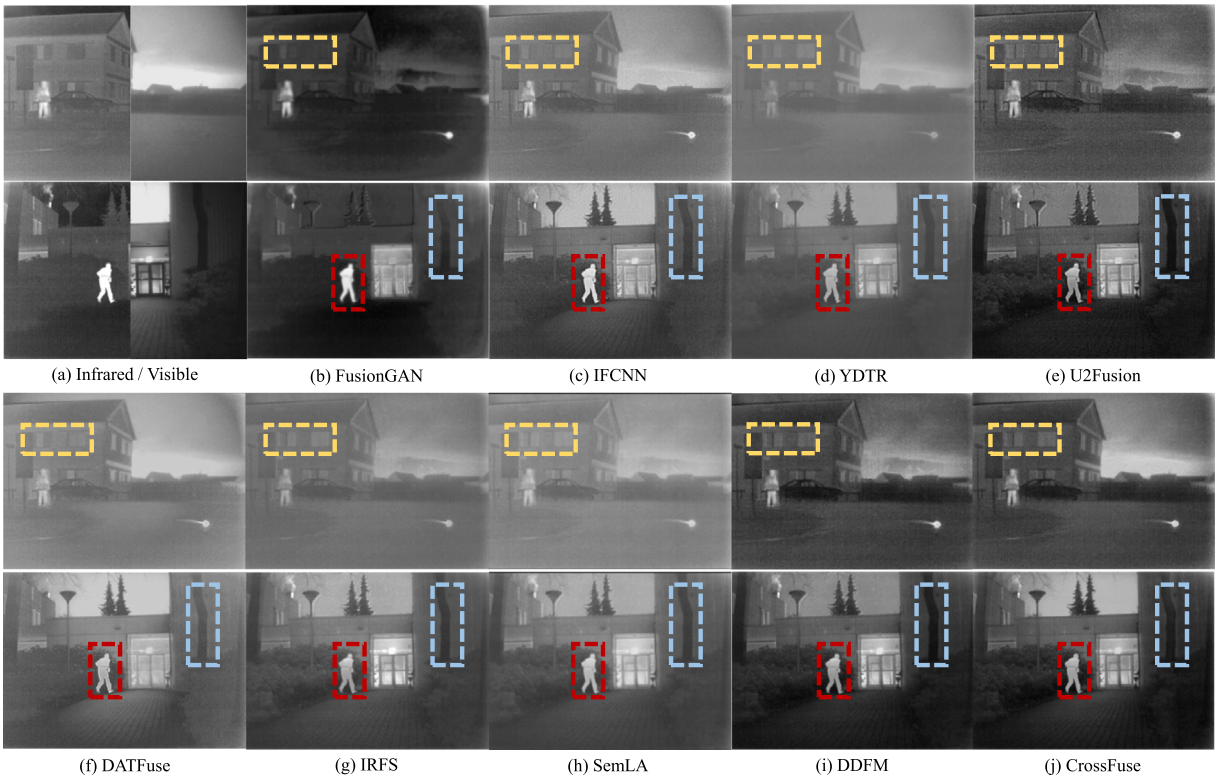

Fig. 15. The fusion results obtained by compared fusion methods and the proposed method on TNO ("yard" and "man"). (a) Infrared and Visible images; (b) FusionGAN; (c) IFCNN; (d) YDTR; (e) U2Fusion; (f) DATFuse; (g) IRFS; (h) SemLA; (i) DDFM; (i) CrossFuse (*the proposed*).

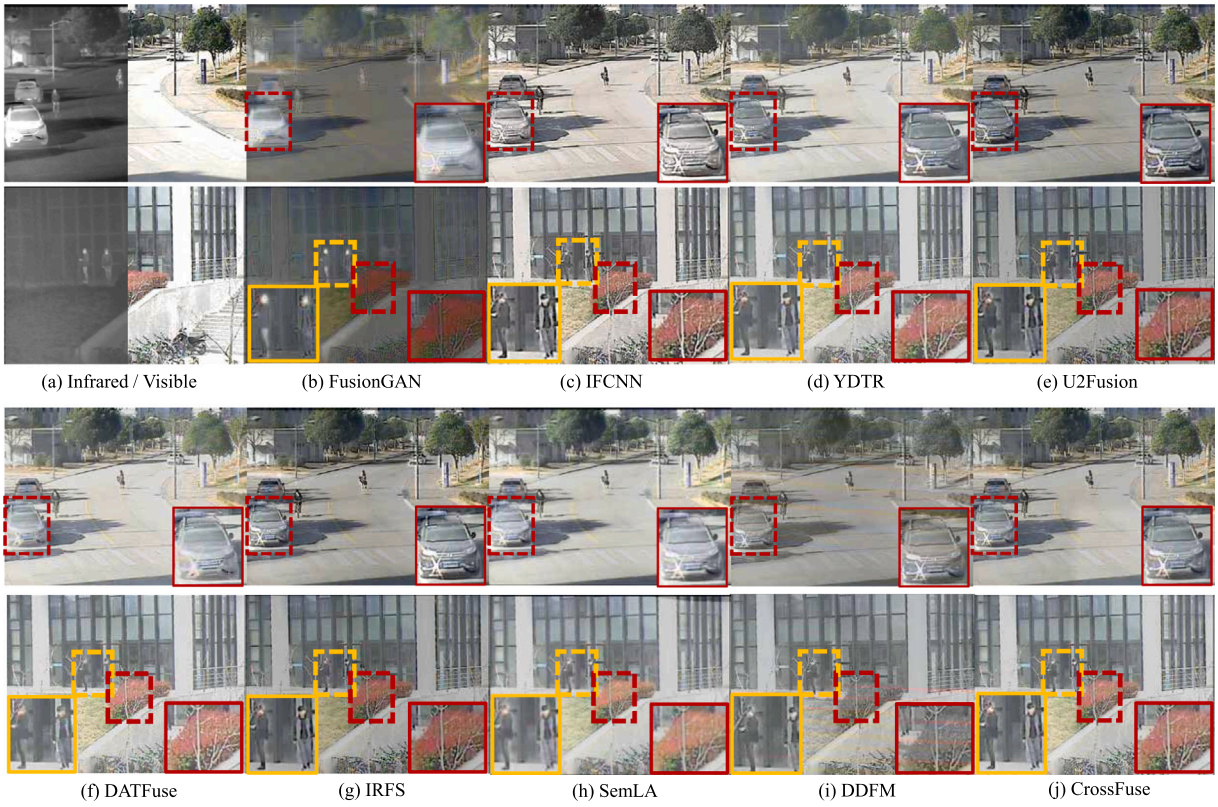

Fig. 16. The fusion results obtained by compared fusion methods and the proposed method on VOT-RGBT ("crossroad" and "two-man"). (a) Infrared and Visible images; (b) FusionGAN; (c) IFCNN; (d) YDTR; (e) U2Fusion; (f) DATFuse; (g) IRFS; (h) SemLA; (i) DDFM; (i) CrossFuse (*the proposed*).

Table 3
The average metrics values obtained by the existing fusion methods and the proposed network on TNO.

	Year	EN↑	SD↑	MI↑	FMI_{dct} ↑	FMI_{pixel} ↑	SCD↑
FusionGAN [21]	2020	6.3629	54.3575	12.7257	0.3634	0.8906	1.4569
IFCNN [32]	2021	6.5955	*66.8758*	13.1909	*0.3738*	0.9009	1.7138
YDTR [34]	2022	6.2268	51.4882	12.4536	0.3483	0.8992	1.6106
U2Fusion [30]	2022	*6.7571*	64.9116	*13.5142*	0.3406	0.8936	**1.7984**
DATFuse [42]	2023	6.3206	56.0363	12.6412	0.2738	0.8807	1.5240
IRFS [35]	2023	6.43326	59.13428	12.86652	0.37997	0.90520	1.74604
SemLA [60]	2023	6.52166	63.92465	13.04331	0.16495	0.90865	1.53009
DDFM [64]	2023	6.72427	66.64661	13.44855	0.21777	0.88257	1.54674
CrossFuse	*ours*	**6.8389**	**73.4712**	**13.6779**	**0.3866**	*0.9044*	*1.7659*

Table 4
The average metrics values obtained by the existing fusion methods and the proposed network on VOT-RGBT.

	Year	EN↑	SD↑	MI↑	FMI_{dct} ↑	FMI_{pixel} ↑	SCD↑
FusionGAN [21]	2020	6.5203	62.8494	13.0406	0.3646	0.8912	1.3748
IFCNN [32]	2021	6.7411	76.2492	13.4821	0.3736	0.9047	1.6686
YDTR [34]	2022	6.4012	62.4483	12.8024	0.3461	0.9051	1.5624
U2Fusion [30]	2022	**6.9487**	*76.7838*	**13.8973**	0.3364	0.8970	**1.7479**
DATFuse [42]	2023	6.4580	62.3694	12.9160	0.2745	0.8843	1.4845
IRFS [35]	2023	6.6071	67.5912	13.2141	*0.3740*	0.9059	*1.7117*
SemLA [60]	2023	6.6757	71.5133	13.3513	0.1618	**0.9074**	1.5475
DDFM [64]	2023	6.8214	73.2632	13.6428	0.1805	0.8772	1.5112
CrossFuse	*ours*	*6.8908*	**77.1780**	*13.7816*	**0.3827**	*0.9061*	1.6635

fused image, "Y" and the infrared image (gray-scale) are fused by the fusion method firstly. Then, "Cr" and "Cb" are directly combined with the fused gray-scale image to generate the final fused image in YCrCb space. Finally, the fused image is converted to RGB space.

In Fig. 16, two pairs of infrared and visible images, "crossroad" and "two-man", are chosen to demonstrate the visual results generated by the existing fusion methods and the proposed method. Comparing with IFCNN [32], YDTR [34], U2Fusion [30] and DATFuse [42], the infrared objects are enhanced by the proposed CrossFuse (Fig. 16, red box and yellow box). This observation indicates that the novel cross attention mechanism can enhance more complementary features between different modalities compared with state-of-the-art fusion methods. Furthermore, comparing with FusionGAN [21], IRFS [35], SemLA [60] and DDFM [64], our proposed network preserves more detail information (Fig. 16 "two-man", yellow box) while obtaining the comparable salient objects.

The average values[1] of six metrics are shown in Table 4, the best values and the second-best values are denoted in **blob** and *italic and red*, respectively. The proposed fusion method, CrossFUse, obtains two best values (SD and FMI_{dct}) and three second-best values (EN, MI, FMI_{pixel}). Although the proposed method dose not achieve all best values, comparing with the state-of-the-art fusion methods, it still achieves comparable metrics values and even better fusion performance in image sharpness (SD). These observations indicate that our proposed method obtains better fusion performance in both visual evaluation and objective evaluation.

5. Conclusions

The uncorrelation (complementary) is the key to multimodal image fusion task, which needs to be paid more attention. Unfortunately, the existing transformer based methods ignore this limitation. Thus, a novel hybrid (CNN and transformer) fusion network (CrossFuse) is introduced, in which a new cross attention mechanism (CAM) is proposed and applied to fusion module. The key part of CAM is $re\text{-}softmax(\cdot)$ operation which is utilized to enhance the complementary features between different modalities and reduce the redundant information. Moreover, a simple yet efficient loss function is also proposed to force the fused image contains more salient features and detail information from source images. The experimental results on public datasets show that our proposed fusion network demonstrates better fusion performance than the start-of-the-art fusion methods.

While the proposed cross-attention mechanism proves simple yet efficient in the image fusion task, it has limitations in significantly enhancing fusion performance within the transformer framework. A possible research direction involves incorporating additional machine learning methods, such as sparse representation and metric learning, to augment the effectiveness of the cross-attention mechanism. Future efforts will be directed towards exploring and implementing these solutions.

CRediT authorship contribution statement

Hui Li: Idea, Writing – original draft, Gives the response to AE and reviewers' comments point by point. **Xiao-Jun Wu:** Reviews the manuscript and the response.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62202205, 62332008, 62020106012), the National Key Research and Development Program of China (2023YFE0116300), and the Fundamental Research Funds for the Central Universities (JUSRP123030).

[1] These metrics are calculated under the gray-scale space, which means the visible images are converted to gray-scale firstly.

References

[1] Y. Liu, L. Wang, J. Cheng, C. Li, X. Chen, Multi-focus image fusion: A survey of the state of the art, Inf. Fusion 64 (2020) 71–91.

[2] X. Zhang, Deep learning-based multi-focus image fusion: A survey and a comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2021) 4819–4838.

[3] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.

[4] G. Vivone, Multispectral and hyperspectral image fusion in remote sensing: A survey, Inf. Fusion 89 (2023) 405–417.

[5] G. Pajares, J.M. De La Cruz, A wavelet-based image fusion tutorial, Pattern Recogn. 37 (9) (2004) 1855–1872.

[6] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.

[7] Y. Liu, X. Chen, R.K. Ward, Z.J. Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.

[8] H. Li, X.-J. Wu, Multi-focus image fusion using dictionary learning and low-rank representation, in: International Conference on Image and Graphics, Springer, Cham, Switzerland, 2017, pp. 675–686.

[9] H. Li, X.-J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in: 2018 24th International Conference on Pattern Recognition, ICPR, IEEE, 2018, pp. 2705–2710.

[10] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.

[11] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, Bayesian fusion for infrared and visible images, Signal Process. 177 (2020) 107734.

[12] W. Tang, F. He, Y. Liu, Y. Duan, MATR: Multimodal medical image fusion via multiscale adaptive transformer, IEEE Trans. Image Process. 31 (2022) 5134–5149.

[13] T. Zhou, Q. Li, H. Lu, Q. Cheng, X. Zhang, GAN review: Models and medical image fusion applications, Inf. Fusion 91 (2023) 134–148.

[14] V. Voronin, M. Zhdanova, N. Gapon, A. Alepko, A. Zelensky, E. Semenishchev, Deep visible and thermal image fusion for enhancement visibility for surveillance application, in: Electro-Optical and Infrared Systems: Technology and Applications XIX, Vol. 12271, SPIE, 2022, pp. 198–203.

[15] G. Yadav, D.K. Yadav, Contrast enhancement of region of interest of backlit image for surveillance systems based on multi-illumination fusion, Image Vis. Comput. 135 (2023) 104693.

[16] Z. Wang, Y. Ma, Y. Zhang, Review of pixel-level remote sensing image fusion based on deep learning, Inf. Fusion (2022).

[17] M. Ma, W. Ma, L. Jiao, X. Liu, L. Li, Z. Feng, S. Yang, et al., A multimodal hyper-fusion transformer for remote sensing image classification, Inf. Fusion 96 (2023) 66–79.

[18] X. Liang, C. Jung, Deep cross spectral stereo matching using multi-spectral image fusion, IEEE Robot. Autom. Lett. 7 (2) (2022) 5373–5380.

[19] L. Liu, X. Song, J. Sun, X. Lyu, L. Li, Y. Liu, L. Zhang, MFF-Net: Towards efficient monocular depth completion with multi-modal feature fusion, IEEE Robot. Autom. Lett. 8 (2) (2023) 920–927.

[20] H. Li, X.-J. Wu, DenseFuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2019) 2614–2623.

[21] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.

[22] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217.

[23] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, L. Van Gool, CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5906–5916.

[24] H. Li, T. Xu, X.-J. Wu, J. Lu, J. Kittler, LRRNet: A novel representation learning guided fusion network for infrared and visible images, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 11040–11052.

[25] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1037–1045.

[26] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5802–5811.

[27] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, P. Li, DIDFuse: Deep image decomposition for infrared and visible image fusion, in: IJCAI, ijcai.org, 2020, pp. 970–976.

[28] H. Li, X.-J. Wu, J. Kittler, RFN-Nest: An end-to-end residual fusion network for infrared and visible images, Inf. Fusion 73 (2021) 72–86.

[29] Z. Wang, Y. Chen, W. Shao, H. Li, L. Zhang, SwinFuse: A residual swin transformer fusion network for infrared and visible images, IEEE Trans. Instrum. Meas. 71 (2022) 1–12.

[30] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 4 (1) (2022) 502–518.

[31] H. Li, X.-J. Wu, T.S. Durrani, Infrared and Visible Image Fusion with ResNet and zero-phase component analysis, Infrared Phys. Technol. 102 (2019) 103039.

[32] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, IFCNN: A general image fusion framework based on convolutional neural network, Inf. Fusion 54 (2020) 99–118.

[33] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, J. Liu, Efficient and model-based infrared and visible image fusion via algorithm unrolling, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2022) 1186–1196.

[34] W. Tang, F. He, Y. Liu, YDTR: Infrared and visible image fusion via y-shape dynamic transformer, IEEE Trans. Multimed. (2022).

[35] D. Wang, J. Liu, R. Liu, X. Fan, An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection, Inf. Fusion 98 (2023) 101828.

[36] V. Vs, J.M.J. Valanarasu, P. Oza, V.M. Patel, Image fusion transformer, in: 2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022, pp. 3566–3570.

[37] J. Zhang, L. Jiao, W. Ma, F. Liu, X. Liu, L. Li, P. Chen, S. Yang, Transformer based conditional GAN for multimodal image fusion, IEEE Trans. Multimed. (2023).

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S.a. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[41] L. Qu, S. Liu, M. Wang, Z. Song, Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 2, 2022, pp. 2126–2134.

[42] W. Tang, F. He, Y. Liu, Y. Duan, T. Si, DATFuse: Infrared and visible image fusion via dual attention transformer, IEEE Trans. Circuits Syst. Video Technol. (2023).

[43] I. Afyouni, Z. Al Aghbari, R.A. Razack, Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey, Inf. Fusion 79 (2022) 279–308.

[44] X.-F. Zhu, T. Xu, X.-J. Wu, Visual object tracking on multi-modal RGB-D videos: A review, 2022, arXiv preprint arXiv:2201.09207.

[45] Y. Li, H. Liu, H. Tang, Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 2, 2022, pp. 1456–1463.

[46] W. Tang, F. He, Y. Liu, TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation, Pattern Recognit. (ISSN: 0031-3203) 137 (2023) 109295.

[47] L. Floridi, M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences, Minds Mach. 30 (2020) 681–694.

[48] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S.W. Zamir, R.M. Anwer, F. Shahbaz Khan, Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, Springer, 2023, pp. 3–20.

[49] F. Yuan, Z. Zhang, Z. Fang, An effective CNN and transformer complementary network for medical image segmentation, Pattern Recognit. 136 (2023) 109228.

[50] Q. Zhou, S. Ye, M. Wen, Z. Huang, M. Ding, X. Zhang, Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer, Neural Comput. Appl. 34 (24) (2022) 21741–21761.

[51] J. Chen, X. Chen, S. Chen, Y. Liu, Y. Rao, Y. Yang, H. Wang, D. Wu, Shape-Former: Bridging CNN and Transformer via ShapeConv for multimodal image matching, Inf. Fusion 91 (2023) 445–457.

[52] A. Jha, S. Bose, B. Banerjee, GAF-Net: Improving the performance of remote sensing image fusion using novel global self and cross attention learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6354–6363.

[53] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (2019) 512–531.

[54] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, Y. Shan, Dual cross-attention learning for fine-grained visual categorization and object re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4692–4702.

[55] R.G. Praveen, W.C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A.L. Koerich, S. Bacon, P. Cardinal, et al., A joint cross-attention model for audio-visual fusion in dimensional emotion recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2486–2495.

[56] B. Kim, H. Jung, K. Sohn, Multi-exposure image fusion using cross-attention mechanism, in: 2022 IEEE International Conference on Consumer Electronics, ICCE, IEEE, 2022, pp. 1–6.

[57] X. Zhou, Z. Jiang, I.P. Okuwobi, CAFNET: Cross-attention fusion network for infrared and low illumination visible-light image, Neural Process. Lett. (2022) 1–15.

[58] Y. Rao, D. Wu, M. Han, T. Wang, Y. Yang, T. Lei, C. Zhou, H. Bai, L. Xing, AT-GAN: A generative adversarial network with attention and transition for infrared and visible image fusion, Inf. Fusion 92 (2023) 336–349.
[59] L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, SuperFusion: A versatile image registration and fusion network with semantic awareness, IEEE/CAA J. Autom. Sin. 9 (12) (2022) 2121–2137.
[60] H. Xie, Y. Zhang, J. Qiu, X. Zhai, X. Liu, Y. Yang, S. Zhao, Y. Luo, J. Zhong, Semantics lead all: Towards unified image registration and fusion from a semantic perspective, Inf. Fusion 98 (2023) 101835.
[61] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, CCNET: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.
[62] A. Toet, TNO image fusion dataset, 2014, https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029.
[63] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L.Č. Zajc, A. Lukežič, O. Drbohlav, et al., The eighth visual object tracking VOT2020 challenge results, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 547–601.
[64] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, L. Van Gool, DDFM: Denoising diffusion model for multi-modality image fusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 8082–8093.
[65] J.W. Roberts, J.A. Van Aardt, F.B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens. 2 (1) (2008) 023522.
[66] Y.-J. Rao, In-fibre Bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355.
[67] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[68] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Comput. Electr. Eng. 37 (5) (2011) 744–756.
[69] V. Aslantas, E. Bendes, A new image quality metric for image fusion: The sum of the correlations of differences, AEU-Int. J. Electron. Commun. 69 (12) (2015) 1890–1896.