Scientific Reports

https://doi.org/10.1038/s41598-026-45802-9

Article in Press

Infrared-visible image fusion with double-attention mechanism and adaptive interaction loss

Received: 14 January 2026

Accepted: 23 March 2026

Published online: 03 April 2026

Cite this article as: Wang Z., Hu Y. & Zhang B. Infrared-visible image fusion with double-attention mechanism and adaptive interaction loss. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-45802-9

Ziqian Wang, Yanxiang Hu & Bo Zhang

Infrared-Visible Image Fusion with Double-Attention Mechanism and Adaptive Interaction Loss

Ziqian Wang[1], **Yanxiang Hu**[2,*], **and Bo Zhang**[3]

[1,2,3]Tianjin Normal University, College of Computer and Information Engineering, Tianjin, 300387, China
[*]huyanxiang@tjnu.edu.cn

ABSTRACT

Infrared and visible light image fusion aims to synthesize a more informative result by extracting and integrating complementary salient features within two heterogeneous modalities. Recent research has shown that capturing explicit self-similarity and implicit cross-correlation with the aid of an attention mechanism has garnered significant interest and presents several advantages. However, exploring the complementary relationships more comprehensively and optimizing the interaction degrees of double attentions quantitatively is still a challenging issue. In this paper, a novel infrared and visible light image fusion method exploring a double-attention mechanism is proposed. Specifically, our approach excavates intra- and inter-attention features of source images through a two-step feature extraction strategy and integrates them with an intra-attention block in the feature fusion stage. Additionally, to regulate the interaction of two kinds of attentions optimally, an adaptive interaction loss term is devised. In these ways, the salient infrared targets and visible texture details can be integrated more effectively. In the experiments, the proposed method was contrasted with seven state-of-the-art methods on the TNO and RoadScene datasets. The comprehensive subjective and objective comparisons demonstrate the superiority of our method. In addition, a thorough experiment and discussion on the interaction of intra- and inter-information is presented to validate and analyze the effectiveness of our work further.

Keywords: Infrared and visible image fusion, cross-modal inter-attention, adaptive training loss, Transformer.

1 Introduction

With the rapid development of imaging technology, the natural scene can be observed from different spectral ranges. These images with different imaging mechanisms provide multiple perspectives about the observed scene. As an important image enhancement technology, Infrared and Visible Light Image Fusion (IVIF) has important value for visual perception and downstream vision tasks, such as image segmentation and object recognition[1,2] , remote sensing[3,4], military security surveillance[5,6], and autonomous driving[7], and salient object detection[8] etc.

Infrared images sense the thermal radiation of imaging scenes, thus they can capture high-contrast salient thermal targets under all-weather conditions, but the low spatial resolutions make them lack enough details. Conversely, visible light images receive the reflected visible light and have high spatial resolutions, thus they exhibit rich texture details but are highly fragile to environmental and atmospheric interferences. Therefore, merging them to generate more informative images is of significant value, and new research findings are emerging continuously[9,10].

In recent years, the self-attention mechanism-based Vision Transformer (ViT)[11] has gained huge popularity in various CV domains[12], and the same applies to image fusion[13–16]. For diverse image generation tasks, its significance lies in inferring that the local regions can draw support from nonlocal dependency. As to IVIF, besides global self-similarity, cross-modal intercorrelation is also an effective way to further assist fusion quality. It has the capability to uncover the latent relevance between infrared and visible light spectral bands effectively. Based on this distinctive advantage, many IVIF methods have tried to exploit the inter-attention between two sources in different ways[13–16]. For example, SwinFuse[16] employs Swin-Transformer (ST) to extract the self-attention features of two sources firstly, and then these self-features are combined through a weighted strategy. Because the image features are extracted separately, the cross-modal interaction degree is limited. Some recent works exploited cross-modality interaction in more coordinated ways. For example, in GAN-based TGFusion[14], the sources are concatenated in the channel dimension and then sent to a CNN-based UNet generator to extract the mixed multi-scale features. In the feature fusion stage, a Transformer module is inserted to explore the long-dependency of these mixed features. A step forward, SwinFusion[13] improves this deficiency by introducing an ST-based cross-modal interaction step in the feature extraction stage. But in the feature fusion stage, they only use a simple CNN layer. This weakens the utilization of these heterogeneous deep features. From the above discussions, exploring a more collaborative network framework suitable for double-attention mechanisms is a worthy direction.

From another perspective, training loss functions also have pivotal effects on fusion quality, especially for the unsupervised IVIF task. The widely used IVIF loss terms include intensity, gradient, Structural Similarity (SSIM), perceptual loss, etc.[3,9]. These loss terms have demonstrated their remarkable effectiveness, but they guide network training from a result-oriented perspective and cannot provide direct fine-grained process control. For double-attention mechanism-based IVIF methods, new loss terms that can regulate the double-attention interaction degrees adaptively, and consequently achieve the goal of integrating all salient information, are of significant research value.

Concluding the above discussion, the double-attention mechanism-based framework has been demonstrated to be an effective IVIF way, but how to excavate the double-attention information more comprehensively, and more importantly, control the interaction of two factors optimally, still has large room to be explored. Motivated by the above analysis, we propose a new IVIF method exploring double-attention in this paper. We aim to design a more effective model to mine the cross-modal dependencies within two heterogeneous source images, and at the same time, more importantly, elaborate on an effective way to achieve the best interaction regulation. Specifically, our main work includes:

1. A new double-attention mechanism-based IVIF method is proposed. Our model includes a two-step double-attention feature extraction stage, a self-attention-based feature fusion stage, and a result reconstruction stage. It achieves the effective excavation and integration of the unique and complementary features with the aid of a double-attention mechanism.

2. An adaptive interaction loss term is devised to regulate the interaction degrees of two types of attention. It adjusts the attention interaction degrees by taking the local saliency of source images into account and therefore, strengthens local salient features. In this way, we can control the contributions of the two attentions to the final fusion features effectively. However, to the best of our knowledge, no existing method has considered the optimal interaction and interaction degree regulation between the two modalities.

3. A large number of subjective and objective experiments and comparisons were conducted on the TNO and RoadScene datasets. The experimental results demonstrate that our method presents competitive advantages. In addition, an in-depth comparison and analysis of the double-attention mechanism is presented based on a large amount of ablation experiments.

In Fig. 1, we show the conclusive Windrose map comparisons between our method and seven SOTA IVIF methods on two datasets. The six assessment metrics include standard deviation (SD), mutual information (MI), Qabf, spatial frequency (SF), entropy (EN), and average gradient (AG). It is obvious that our method presents overall advantages, especially on SD. This means our fusion results present larger intensity contrast degrees and more details.

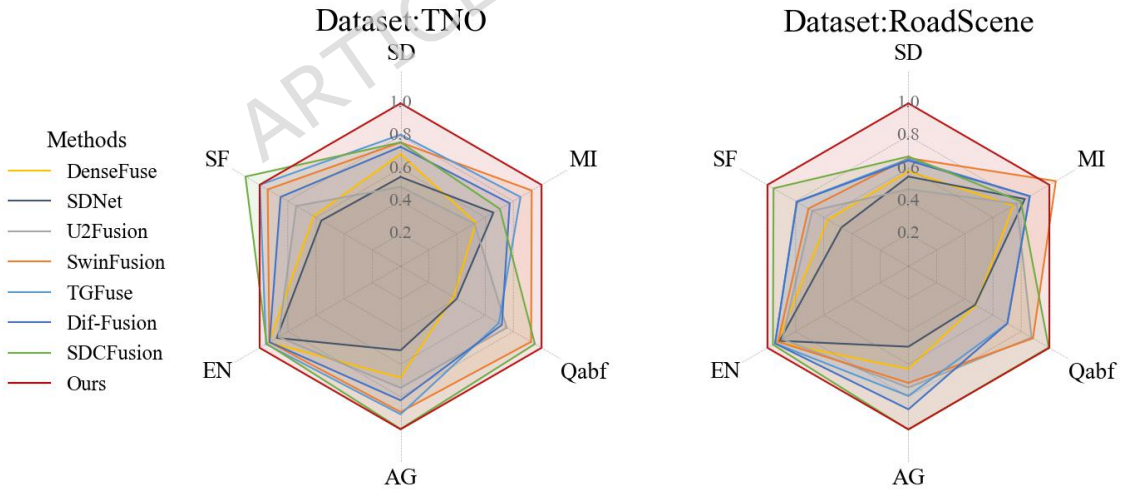


Figure 1. The Windrose map comparisons of our method and seven SOTA methods on two IVIF datasets.

The remainder of the paper is organized as follows: in Section 2, the review of deep learning based IVIF methods and ViT is presented; then, in Section 3, the proposed method is introduced in detail; next, the comprehensive subjective and objective experimental results are reported and discussed in Section 4; and finally, the paper is concluded in Section 5.

2 Related work

2.1 Deep learning based IVIF methods

In the IVIF domain, deep learning based methods have already taken an overwhelming position. These methods are mainly divided into four categories: Auto-Encoder (AE) based methods, CNN-based methods, GAN-based methods, and Transformer-based methods[3,10,17].

AE-based methods consist of three separate stages: feature extraction, feature fusion, and fusion result reconstruction. A pretrained AE is responsible for extracting the features of infrared and visible light images; then these extracted features are combined with some handcrafted rules; finally, a pretrained decoder reconstructs the fusion images. For example, DenseFuse[18] employs a CNN encoder with dense-connection to extract the features of source images, then these features are combined with addition or L_1 norm fusion rules and sent to the decoder for result reconstruction. For this type of method, the research focuses on fusion rule optimization and network architecture improvement. To compensate for the limitations of handcrafted fusion rules, Li[19] and Zhang[20] employed a saliency degree-based adaptive strategy for feature fusion. In addition, some methods used multiple encoders or multiple decoders to alleviate the vital information loss[21,22].

In general, CNN-based methods are end-to-end, i.e., they don't depend on the explicit manual fusion rules. Because of the unsupervised learning essence of IVIF tasks, the targeted loss function design is crucial. Besides loss functions, the network architectures are another central issue. Liu et al.[23] proposed one of the first CNN-based IVIF methods. They employed a Siamese CNN to obtain a weight map from source images at first and then stitched them to generate a fusion result. Following their work, many CNN-based IVIF methods have been proposed based on different network structures and learning strategies, such as UNet[24], DenseNet[18,25], ResNet[26], dilated convolution[27], multiscale feature[28,29], and hybrid architecture[30], etc.

GAN relies on adversarial games between the generator and the discriminator to model the implicit source data distribution. Since Ma et al.'s pioneering work[31], many GAN-based image fusion methods have been proposed[32]. For IVIF, how to balance the adversarial loss of infrared and visible light images is a key issue. Some early IVIF methods employ a discriminator to evaluate the confidence levels of the fusion results[33,34]. However, they tend to bias the fusion results towards one of the source images. Some improved methods harness dual discriminators to measure the deviation degrees between the fusion results and one of two source images, respectively[35,36], which makes the fusion results contain more thermal targets and textures simultaneously. For instance, DDcGAN[36] uses two discriminators and a conditional GAN to improve its fusion performance.

Compared with convolution operations, the self-attention mechanism-based Transformer presents strong global context modeling and representation capability. In image fusion tasks, global dependency can help to infer the local fusion results according to the non-local long-range similarity. In practice, Transformer is usually combined with CNN to construct the feature extraction module, i.e., the Transformer block excavates the long-range dependences on the basis of the shallow CNN features. In addition, some Transformer-based IVIF methods employ a Transformer in the feature fusion stage. For example, based on the features extracted with the multiple-layer CNN, SwinFusion[13] used an intra-domain ST and an inter-domain ST for feature fusion. In[14], Rao et al. employed a spatial Transformer and a cross-channel Transformer for bidirectional global dependency modeling in a GAN framework. With the aid of double discriminators, the integration performance of the infrared and visible light modals is improved. Unlike the methods mentioned above, some other methods apply a Transformer in the feature extraction stage. For example, SwinFuse[16] employed ST blocks with residual connections to extract intra-modal global features. ADFNet[37] structurally extracts low-frequency global features via Transformer and high-frequency local features via CNN, fuses the two types of features to generate images. Similar work also includes[38,39] et al. In addition, DGLT-Fusion[40] and TCCFusion[15] employ independent Transformer-based global and local feature extraction modules for further quality improvement.

2.2 Swin-Transformer

In 2020, ViT[11] was proposed for image and video processing. ViT developed a self-attention mechanism to mine the dependencies between the image patches and uses these similarities for feature representation. A significant drawback of ViT is that the computational cost increases with the square ratio of image resolution, and this seriously limits its practicality. By computing the self-attention weights within a window and shifting the window regularly, ST[41] decreases the computational cost evidently and promotes processing efficiency. By now, ST has become the mainstream computational way of self-attention and has been utilized in many CV tasks[42–44].

ST captures cross-window self-attention by combining with window-based multi-head self-attention (W-MSA) and shifted-window-based multi-head self-attention (SW-MSA) strategies. The detailed process can be formulated as:

$$\hat{\mathscr{Z}}^{l} = \text{W-MSA}\left(\text{LN}\left(\mathscr{Z}^{l-1}\right)\right) + \mathscr{Z}^{l-1} \tag{1}$$

$$\mathscr{Z}^{l} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathscr{Z}}^{l}\right)\right) + \hat{\mathscr{Z}}^{l} \tag{2}$$

$$\hat{\mathscr{Z}}^{l+1} = \text{SW-MSA}\left(\mathrm{LN}\left(\mathscr{Z}^{l}\right)\right) + \mathscr{Z}^{l} \tag{3}$$

$$\mathscr{Z}^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathscr{Z}}^{l+1}\right)\right) + \hat{\mathscr{Z}}^{l+1} \tag{4}$$

where LN denotes LayerNorm layer, MLP denotes multilayer perceptron, $\hat{\mathscr{Z}}^{l}$is the output of W-MSA block, $\hat{\mathscr{Z}}^{l+1}$ is the output of SW-MSA block, $\mathscr{Z}^{l}$ and $\mathscr{Z}^{l+1}$ are the output of MLP, more detailed explanations can be found in[41].

On the basis of image patching, W-MSA partitions the image patches into N*N no-overlap local windows. Then, self-attention is computed within each local window:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \tag{5}$$

where Q, K, and V denote query, key, and value components of the patches, d is the dimension of Q and K, and B is the position bias.

To expand the attention range, SW-MSA employs elaborated windows partitioning and shifting rules to cover the whole image alternately. As shown in Fig. 2, on the basis of regular window partitioning in the l-th layer, SW-MSA shifts the window up and left by $\lfloor M/2, M/2 \rfloor$ pixels to construct the l+1-th layer partitioning and then carries out attention computation within the new windows.

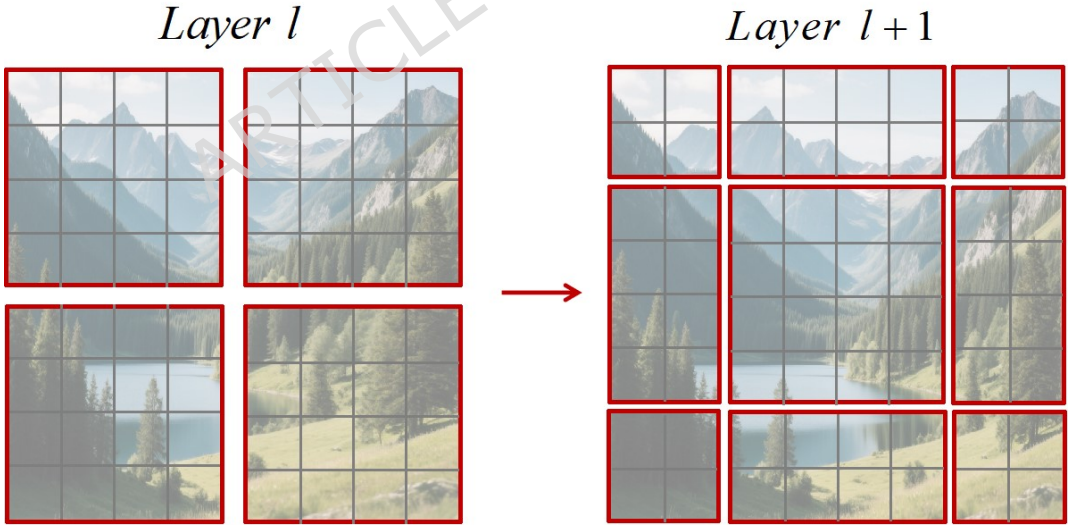


Figure 2. The illustration of the shifted window approach in ST.

3 The proposed method

In this section, we introduce our method in detail. Concretely, the overall network architecture and fusion process are presented in Section 3.1; in Section 3.2, the double-attention mechanism-based feature extraction stage is detailed; and in Section 3.3, the loss function is introduced.

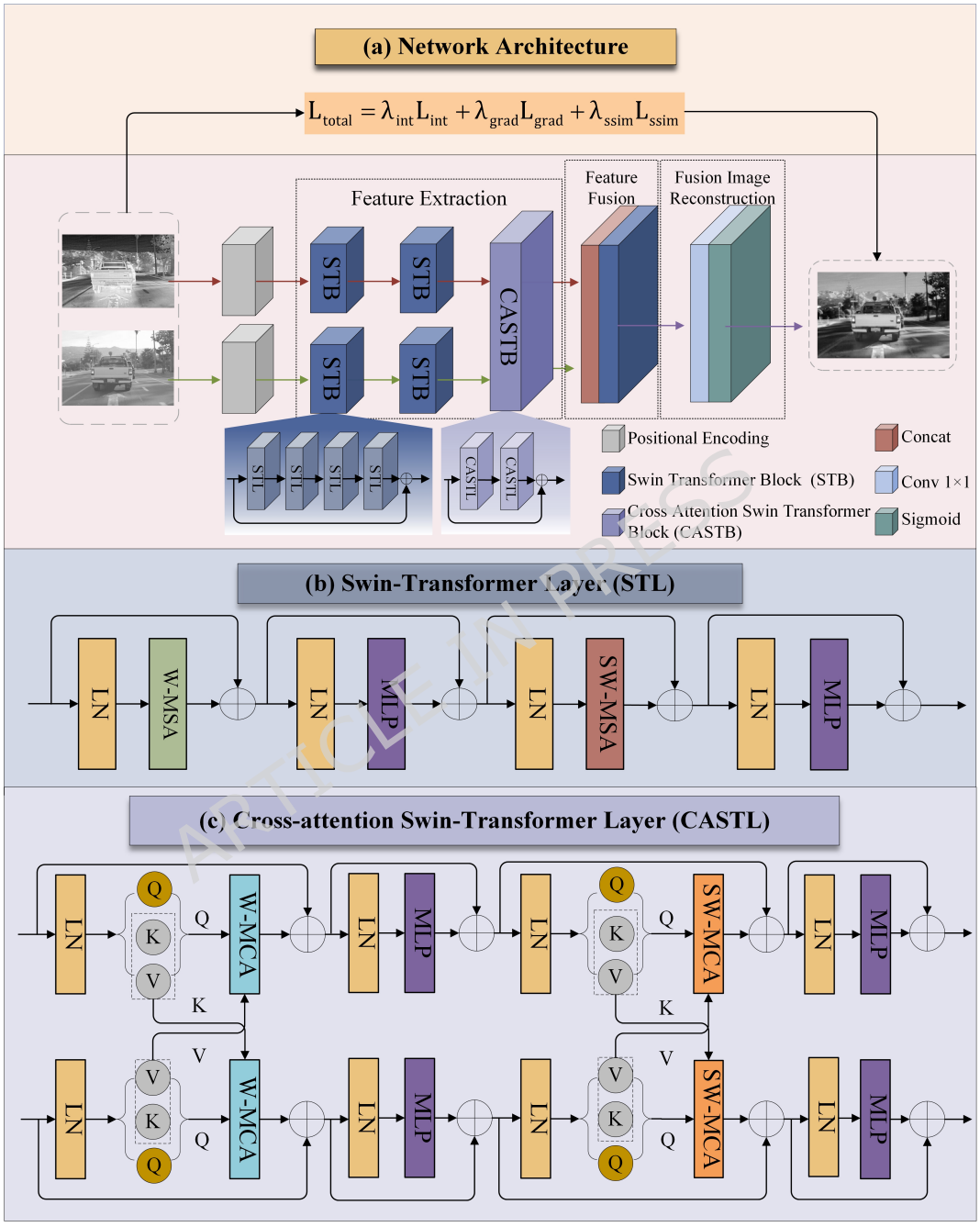


Figure 3. The overall network architecture of our model.

3.1 Network architecture and fusion process

On the whole, our method is an end-to-end model that generates the fusion image with a feedforward pretrained network. Fig. 3 shows the overall architecture.

As shown in Fig. 3(a), our model includes three stages: source feature extraction, feature fusion, and fusion image reconstruction. The first stage consists of two steps: intra-attention computation and inter-attention interaction. The former includes two Swin-Transformer Blocks (STBs), and the latter includes a Cross-attention Swin-Transformer Block (CASTB); their organizations are shown in Fig. 3(b) and Fig. 3(c). Specifically, each STB contains four Swin-Transformer layers (STL), and each CASTB contains two cross-attention Swin-Transformer layers (CASTL). The second stage receives the dual-way feature representations of two sources and fuses them with a concatenation block and an STB block. The final reconstruction stage includes a 1*1 convolution layer and a sigmoid layer.

Suppose the sizes of the input images are $I_{i/v} \in \mathbb{R}^{H\times W\times C_{in}}(C_{in}=1)$. The total processing is detailed as follows:

1. Feature extraction stage

(1). Patching and position embedding: the input images are segmented into 16*16 patches at first, and then the patches are positional embedded with a convolutional block:

$$\phi_i^0 = \text{PatchEmbed}(I_i) \tag{6}$$

$$\phi_v^0 = \text{PatchEmbed}(I_v) \tag{7}$$

where i and v denote infrared and visible light features, PatchEmbed denotes patching and positional embedding. The number of output channels C_{out} is set to 1.

(2). After positional embedding, the obtained infrared feature ϕ_i^0 and ϕ_v^0 visible feature are sent into the two-step feature extraction stage, which contains two STBs and one CASTB:

$$\begin{cases} \phi_i^l = \text{STB}\left(\phi_i^{l-1}\right), & l=1,2 \\ \phi_i^3 = \text{CASTB}\left(\phi_i^2, \phi_i^2\right) \end{cases} \tag{8}$$

$$\begin{cases} \phi_v^l = \text{STB}\left(\phi_v^{l-1}\right), & l=1,2 \\ \phi_v^3 = \text{CASTB}\left(\phi_v^2, \phi_v^2\right) \end{cases} \tag{9}$$

here the STB and CASTB are responsible for the intra-attention and cross-attention computation.

2. Feature fusion stage

This stage merges the interacted dual-way source features and generates the fused features for result reconstruction. In this stage, we first concatenate the input features as a whole and then use an STB to mix them. As contrasted with Fig. 4(a), (b), and (c), three widely used feature fusion strategies in the existing IVIF methods are direct concatenation and convolution after concatenation. This simple processing is not conducive to fully taking advantage of cross-modality and long-range interaction.

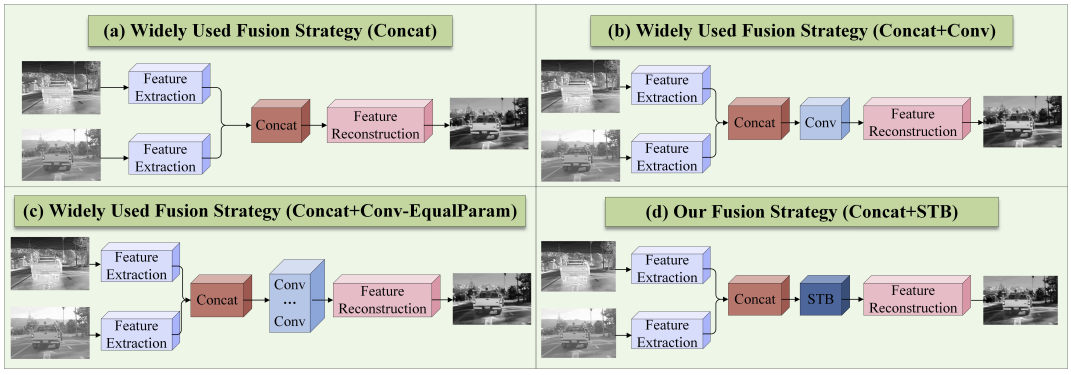


Figure 4. The contrast of four feature fusion strategies.

In contrast, our method utilizes a self-attention way to connect these deep features from a holistic viewpoint, thus it can mix two attentions more deeply. Our processing can be denoted as:

$$\phi_f = \text{Concat}\left(\phi_i^3, \phi_v^3\right) \tag{10}$$

$$\psi_f = \text{STB}\left(\phi_f\right) \tag{11}$$

where ϕ_i^3 and ϕ_v^3 denote the infrared and visible light features output from the CASTB, Concat. denotes concatenation operation and ψ_f is the fused feature after the STB fusion processing.

3. Fusion image reconstruction

In this stage, the fused features ψ_f are sent into the reconstruction module which includes a 1*1 convolutional layer and a Sigmoid activation layer to reconstruct the fusion result I_f:

$$I_f = \text{sigmoid}\left(\text{Conv}_{1\times1}\left(\psi_f\right)\right) \tag{12}$$

4. Training loss. In the training, except for the classical intensity loss and gradient loss, we devise a new adaptive interaction loss term to regulate the double-attention interaction. The details will be introduced in Section 3.3.

3.2 Double-attention feature extraction

In our model, the second step of the feature extraction stage computes the inter-attention between the two source images. As shown in Fig. 3(c), the designed CASTB contains two cross-attention ST layers (CASTL) for deep interaction, and each CASTL includes a cross-modal W-MSA (W-MCA) and a cross-modal SW-MSA (SW-MCA). Suppose the self-attention patch features extracted by the front-end STBs are X_i and X_v, the principle of the W-MCA and SW-MCA can be formulated as:

$$\{Q_i, K_i, V_i\} = \{X_i W_i^Q, X_i W_i^K, X_i W_i^V\} \tag{13}$$

$$\{Q_v, K_v, V_v\} = \{X_v W_v^Q, X_v W_v^K, X_v W_v^V\} \tag{14}$$

$$\text{CrossAttention}(Q_i, K_v, V_v) = \text{Softmax}\left(\frac{Q_i K_v^T}{\sqrt{d}} + B\right) V_v \tag{15}$$

$$\text{CrossAttention}(Q_v, K_i, V_i) = \text{Softmax}\left(\frac{Q_v K_i^T}{\sqrt{d}} + B\right) V_i \tag{16}$$

In equation (13) and equation (14), the preceding features X_i and X_v are transformed to the Q/K/V space at first, and in equation (15) and equation (16), Q_i and Q_v are used to compute their similarities with K_v and K_i. According to the two similarities, the cross-attention weights are obtained. Taking infrared image features as an example, the detail transformations are as follows:

$$\hat{\mathscr{Z}}_i^l = \text{W-MCA}\left(\text{LN}\left(\mathscr{Z}_i^{l-1}\right)\right) + \mathscr{Z}_i^{l-1} \tag{17}$$

$$\mathscr{Z}_i^l = \text{MLP}\left(\text{LN}\left(\hat{\mathscr{Z}}_i^l\right)\right) + \hat{\mathscr{Z}}_i^l \tag{18}$$

$$\hat{\mathscr{Z}}_i^{l+1} = \text{SW-MCA}\left(\text{LN}\left(\mathscr{Z}_i^l\right)\right) + \mathscr{Z}_i^l \tag{19}$$

$$\mathscr{Z}_i^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\mathscr{Z}}_i^l\right)\right) + \hat{\mathscr{Z}}_i^{l+1} \tag{20}$$

where the variable and function representations are similar to equation (1)– equation (4).

For the whole image, the computation process of the window-based cross-attention is shown in Fig. 5. Similar cross-modal mechanisms have also been widely applied in related multi-modal tasks[45,46].

As illustrated in Fig. 5, in layer l (left), cross-modal inter-attention computation is performed under regular window partitioning. In layer l+1 (right), the shifted window partitioning forms new windows, where cross-modal inter-attention computations transcend the boundaries of previous windows, enabling feature interaction across windows.

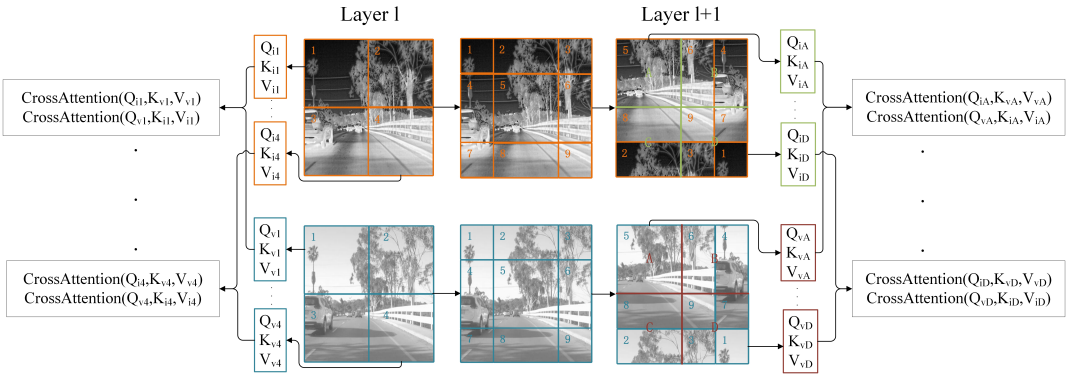


Figure 5. The illustration of the cross-modal shifted window attention.

3.3 Loss function

In IVIF, different prior based training loss terms, such as intensity, gradient, SSIM, and perceptional loss etc., are employed jointly to force the salient information to be transferred to the fusion images. Between them, the intensity loss term constrains the fusion image to keep more infrared target contours and visible background distributions, the gradient loss term forces more detail changes to be collected, and the perceptional loss term uses the convolutional feature maps extracted by VGGNet16 to retain the deep feature similarity. In addition, SSIM is another widely used loss measurement. SSIM[47] compares the similarity of two image patches from three aspects: intensity, contrast degree, and covariance. Thus, SSIM loss emphasizes the regional semantic consistency more. In the specific implementation, the existing IVIF methods all assume that the two sources have equal weights:

$$\mathscr{L}_{\text{SSIM}} = \lambda_1 \cdot \text{SSIM}(I_f, I_v) + \lambda_2 \cdot \text{SSIM}(I_f, I_i) \tag{21}$$

where λ_1 =λ_2 =0.5.

In fact, the setting of equal weights is ill-considered due to the fact that two sources have different saliency in different locals. From the perspective of inter-attention, equal loss weights mean intra-attention and inter-attention have the same influences to the final results. This inevitably leads to the suppression of the intra-modality saliency information.

In order to regulate the contribution degrees of the two kinds of attention adaptively to assimilate all the source significant information, we devised an adaptive SSIM-based loss term to guide the collaboration of the double-attention mechanism. For a region which contains more salient infrared targets, this loss term should highlight infrared attention moderately; conversely, it would emphasize the attention of visible light images with a reasonable ratio. In this way, the network is guided to learn the optimal double-attention interaction degrees.

Concretely, two source images and their fusion result are divided into 8*8 small patches at first; and then the SSIM similarities between two source patches and fusion result patches are computed pairwise; finally, the whole interaction loss is computed according to the local saliency of source patches. This process is formulated as follows:

$$\mathscr{L}_{\text{SSIM}} = 1 - \frac{\sum_{m=1}^{N} w_{v,m} \cdot \text{SSIM}(I_{f,m}, I_{v,m}) + w_{i,m} \cdot \text{SSIM}(I_{f,m}, I_{i,m})}{N} \tag{22}$$

where N is the number of 8*8 patches in the images,$I_{f,m}$,$I_{v,m}$ and $I_{i,m}$ are the m-th patches of I_f, I_v and I_i respectively, $w_{v,m}$ and $w_{i,m}$ are the weights between $I_{f,m}$,$I_{v,m}$ and $I_{i,m}$:

$$w_{i,m} = \frac{\|\nabla I_{i,m}\|_1}{\|\nabla I_{i,m}\|_1 + \|\nabla I_{v,m}\|_1} \tag{23}$$

$$w_{v,m} = \frac{\|\nabla I_{v,m}\|_1}{\|\nabla I_{i,m}\|_1 + \|\nabla I_{v,m}\|_1} \tag{24}$$

where ∇ is the Sobel operator and $\|\cdot\|_1$ denotes L_1 norm.

This adaptive regulation mechanism, based on local saliency contrast, enables the network to dynamically adjust the intensity and direction of cross-modal interaction according to the modal characteristics of different regions, thereby achieving more optimal whole integration.

Although the gradient-based adaptive weights defined in equation (23)– equation (24) may be sensitive to high-frequency noise under low-light conditions, the proposed method mitigates this issue through patch-based SSIM computation and gradient summation, which inherently suppresses the influence of isolated noise pixels[48]. Depending on this adaptive regulation mechanism, our model is capable of achieving a more optimal whole integration.

Besides the interaction loss term, an intensity and a gradient loss term are employed jointly to constitute the total training loss. The intensity loss term L_{Int} is used to constrain the intensity distribution consistency between the sources and fusion image. Owing to the fact that the thermal salient targets exhibit high brightness, we force the fusion image to integrate more high intensity:

$$\mathscr{L}_{\text{Int}} = \frac{1}{HW}\left\|I_f - \max(I_i, I_v)\right\|_1 \tag{25}$$

The gradient loss term L_{Grad} is used to transfer the edge and detail information into the fusion image as much as possible. We use the Sobel operator to compute the gradients of the source images, and force the fusion image to present high contrast with the selecting-max strategy:

$$\mathscr{L}_{\text{Grad}} = \frac{1}{HW}\left\|\nabla I_f - \max(\nabla I_i, \nabla I_v)\right\|_1 \tag{26}$$

where ∇ denotes the Sobel operator.

The total loss in our method is summarized as:

$$\mathscr{L}_{\text{Total}} = \lambda_{\text{Int}}\mathscr{L}_{\text{Int}} + \lambda_{\text{Grad}}\mathscr{L}_{\text{Grad}} + \lambda_{\text{SSIM}}\mathscr{L}_{\text{SSIM}} \tag{27}$$

where λ_{Int} , λ_{Grad} and λ_{SSIM} are the hyper parameters. In the training phase, we set them to 10, 10, and 15 according to the experimental comparisons.

3.4 Code availability

Available from the corresponding author on reasonable request.

4 Experiments and discussions

In this section, we report the experiments in detail and discuss some related issues in depth. Concretely, in Section 4.1, we introduce the experimental settings and implementation details; then, we present the subjective and objective experimental results comprehensively in Section 4.2, along with a computational efficiency comparison of different methods; next, we introduce the ablation experiments in Section 4.3; finally, we discuss and analyze some related issues in depth in Section 4.4.

4.1 Implementation details

4.1.1 Experimental environment

Our method was implemented with the PyTorch platform. The acceleration device is the NVIDIA GeForce RTX 4090 GPU. The optimization algorithm is the Adam optimizer. The initial learning rate is set to 1×10^5, and the batch size and training epoch are set to 4 and 10, respectively.

MSRS dataset[49] is employed in the training phase, which contains 1083 pairs of infrared and visible light images. These image pairs cover a variety of complex scenes, such as daytime and nighttime driving and urban scenes. This strengthens the generalization capability effectively. In the training, each source image is divided into 128*128 patches and the corresponding patch pairs are used as training samples.

4.1.2 Experimental setting

Seven deep learning based SOTA methods are selected to verify the performance of the proposed method, including five specific IVIF methods and two general image fusion methods. Five specific methods include DenseFuse[18], FusionGAN[31], TGFuse[14], Dif-Fusion[50], and SDCFusion[51]. They belong to AE, GAN, Transformer, and diffusion architectures. Two general methods are U2Fusion[24] and SwinFusion[13]. Among them, U2Fusion is an end-to-end CNN model. And as reviewed in Section 2, SwinFusion is a Swin-Transformer model.

DenseFuse is an AE-based, specific end-to-end IVIF method, characterized by its dense connection CNN architecture. By adding dense connections in the encoder, more image details are preserved. FusionGAN is a representative GAN based IVIF-specific method, which contains a CNN based generator and a CNN based discriminator. By the adversarial games between the generator and the discriminator and defining an adversarial loss, the generator learns a targeted joint distribution of two source images and constructs a high-quality fusion result. U2Fusion is a versatile method that harnesses UNet as the backbone. By employing the symmetric encoder-decoder architecture and skipping connections, the multi-scale feature extraction and combination are achieved. Dif-Fusion formulates image fusion as a conditional generation problem within a denoising diffusion probabilistic framework, iteratively refining the fused result from random noise under the guidance of source images. SDCFusion is a semantic-driven coupled network that integrates the fusion network and segmentation network to share cross-modality coupled features, achieving a balance between pixel-level detail preservation and semantic-level representation.

We implemented these methods with the source codes opened by the authors and employed their default settings.

In the experiments, we used TNO[52] and RoadScene[53] datasets. The TNO image pairs were captured in outdoor scenes and contain prominent thermal targets and rich textures. RoadScene dataset covers a variety of traffic scenes and contains a lot of moving thermal objects and rich textured backgrounds. These two test datasets have noticeable visual differences and less relevance to the training dataset used. Therefore, using MSRS for training and using TNO/RoadScene for testing constitutes a reliable cross-dataset evaluation.

In the experiments, we employed four self-information-based and two mutual-information-based metrics for objective comparison. These metrics evaluate the fusion quality from various viewpoints. Standard Deviation (SD) measures the variation degree of an image. An image with a high SD value indicates it contains more contrasts and details. Average gradient (AG) computes the average gradient values of all pixels; it indicates the richness of edge details intuitively. Information theory based Entropy (EN) calculates the information uncertainty of the image; the larger the uncertainty, means high information content. Spatial frequency (SF) computes the horizontal and vertical change frequencies in the image; a great change frequency signifies rich details. Two mutual-information based metrics measure the visual contents in the fusion results that are transferred from the sources. Mutual information (MI)[54] measures the dependence degrees between fusion results and source images by computing the mutual information entropy. Qabf[55] measures the amount of edge information transferred from the source images to the fused image. In the experiments, we employed the open codes provided by Liu et al[56] for standardization.

4.2 Experimental results

In this section, we present the subjective and objective experimental results and compare the proposed method with the contrast methods. Concretely, we report the experimental results of the TNO dataset first, and then present the experimental results of the RoadScene dataset. Additionally, a computational efficiency comparison is conducted to evaluate the practicality of different methods.

4.2.1 The experimental results on TNO dataset

Figure 6 shows twenty pairs of TNO source images used in this group of experiments.

(a)IR (b)VIS

Figure 6. Twenty pairs of source images selected from TNO dataset.

For subjective illustration, Fig. 7 and Fig. 8 show two examples, respectively.

As shown in Fig. 7, the infrared image contains a prominent body target, while the visible image contains a large amount of texture details. Comparing the eight fusion images, we can note that DenseFuse, FusionGAN, and U2Fusion preserve the salient information effectively, but the low contrast degrees result in some detail losses; in addition, some visible noise appears. In contrast, our proposed method, SwinFusion and TGFuse, achieve better visual effects. They not only show complete salient thermal targets and rich details, but also exhibit natural transitions between different regions. Compared to SwinFusion,

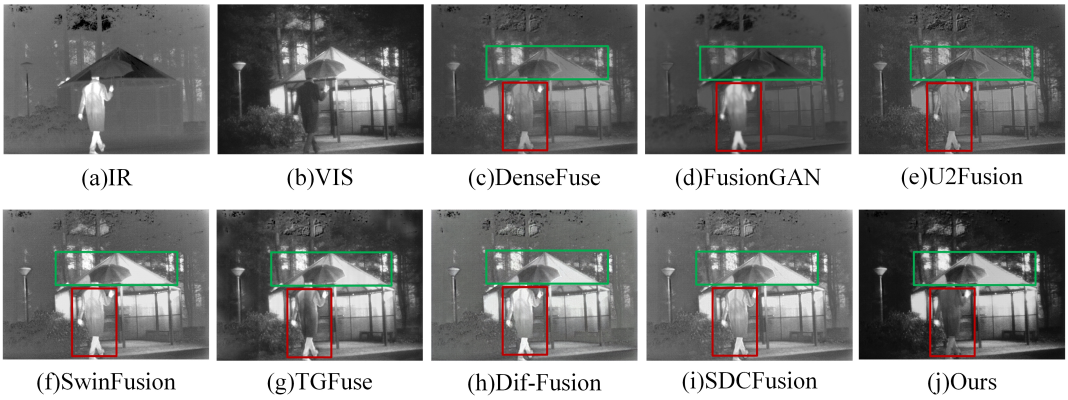


Figure 7. The fusion results of 'Kaptein_1654' image pair.

Dif-Fusion, and SDCFusion, our method demonstrates superior performance in edge clarity and detail richness. As to TGFuse, our approach is superior in terms of detail richness and target saliency. Figure 8 presents another example.

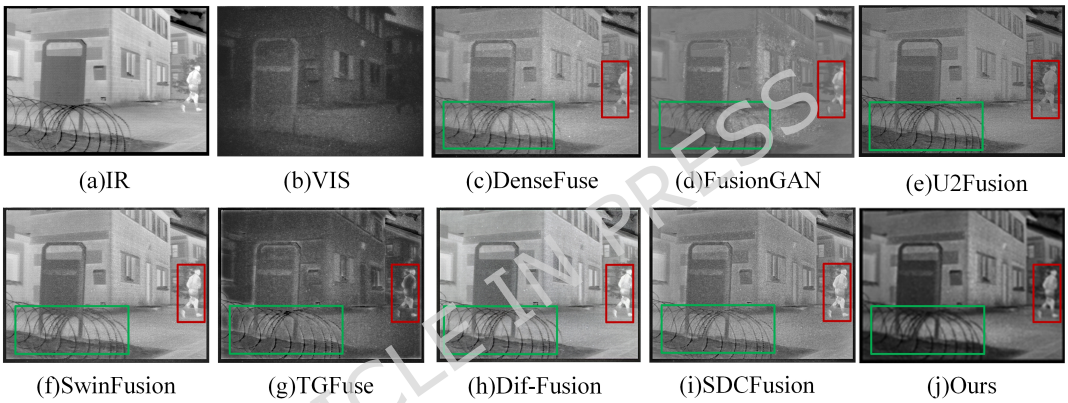


Figure 8. The fusion results of 'barbed_wire' image pair.

As can be seen in Fig. 8, in this group of experiments, the fusion images of DenseFuse, FusionGAN, and TGFuse show low contrast and visual noises; U2Fusion presents high contrast degrees, but the visible light details are lost obviously, such as in the billboards and wire meshes. SwinFusion, Dif-Fusion, SDCFusion, and our method exhibit clear advantages, including clear targets, rich details, and well-situated contrast. Nevertheless, SwinFusion, Dif-Fusion, and SDCFusion are more inclined towards infrared content, thus appear brighter in the whole. Comparatively, our result presents more palatable visual effects and achieves better multimodal information fusion.

Figure 9 shows the objective comparison results of six metrics in this group of tests. The horizontal ordinates denote the twenty examples, and the vertical ordinates denote the six metric values.

As can be seen in Fig. 9, for the six metrics, our method presents significant advantages in SD, Qabf, AG, EN, and SF. These results mean our results contain more contrast and detail changes. As to mutual-information-based MI, our method and SwinFusion present better performance than others. This demonstrates that the attention mechanism can assist the fusion methods in effectively inheriting complementary information from two modalities.

4.2.2 The experimental results on RoadScene dataset

In this experiment, twenty pairs of images from RoadScene datasets were used, as shown in Fig. 10.

For subjective comparison, Fig. 11 and Fig. 12 show two examples respectively.

As shown in Fig. 11, the infrared image contains a prominent car target, while the source visible image contains clear trees nearby and blurry mountains in the distance. Some blurs and artifacts exist in the results of FusionGAN and U2Fusion, such as in the car's license plates and the road signs. Although the results of DenseFuse improve the clarity of the targets, the imbalance between the two modalities leads to some detail loss. The fusion result of SwinFusion has some drawbacks, such as over-brightness, low contrast, and unclear target contours. In addition, due to the Transformer-GAN hybrid framework used,

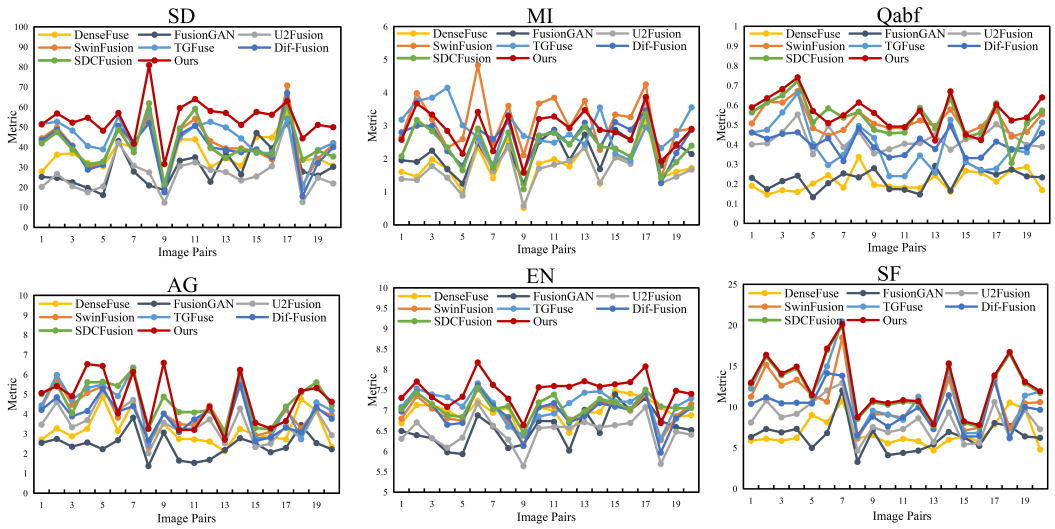
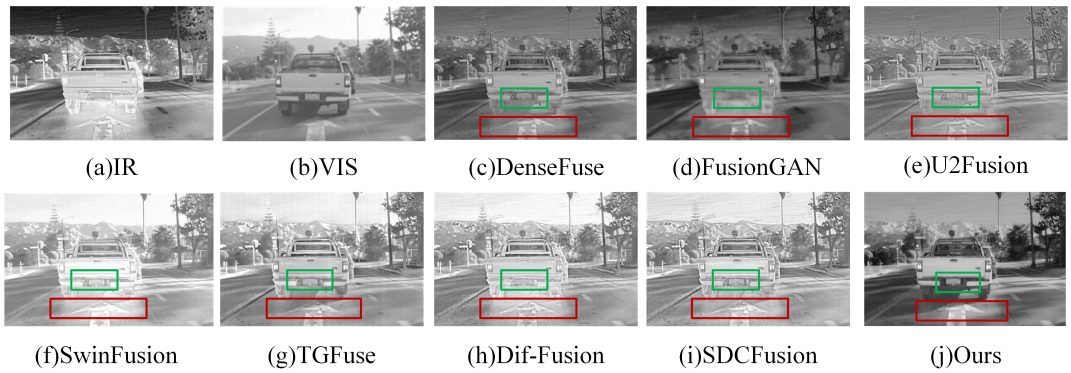


Figure 9. Comparative results of objective metrics for different fusion methods with the TNO dataset.

(a) IR (b) VIS

Figure 10. Twenty pairs of source images selected from RoadScene dataset.



(a)IR (b)VIS (c)DenseFuse (d)FusionGAN (e)U2Fusion

(f)SwinFusion (g)TGFuse (h)Dif-Fusion (i)SDCFusion (j)Ours

Figure 11. The fusion results of 'FLIR_04514' image pair.

TGFuse shows some visual naturalness details, including slightly insufficient sharpness of the car contours and richness of the tree textures. Both Dif-Fusion and SDCFusion tend to retain limited information from the visible image. Their fusion outputs are heavily biased toward the infrared modality, resulting in suppressed background details and reduced texture richness in regions. Figure 12 gives another example for subjective comparisons.

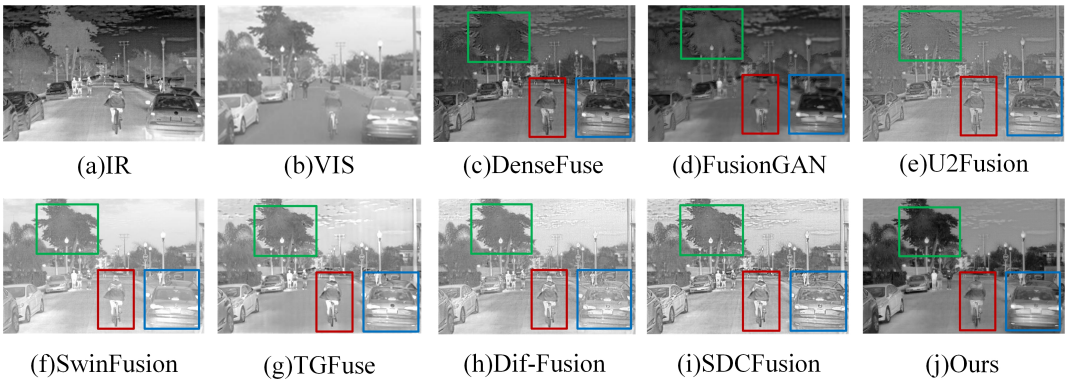


Figure 12. The fusion results of 'FLIR_06832' image pair.

From Fig. 12, we can note that the source images contain multiple significant objects and complex backgrounds. Like in the previous example, the fusion results of FusionGAN, U2Fusion, and SwinFusion present low contrast degrees and clarity, such as the blurred contour of the rider and the background. The modal imbalance in DenseFuse leads to the loss of details, such as the contexts of the rider, bicycle, and trees. Similarly, TGFuse, Dif-Fusion, and SDCFusion present limited detail richness and edge clarity. These comparisons demonstrate our method shows better salient information preservation capability, including all details and contours of the car, bicycle, rider, and tree background.

Figure 13 shows the objective comparison results of six metrics of this group of experiments.

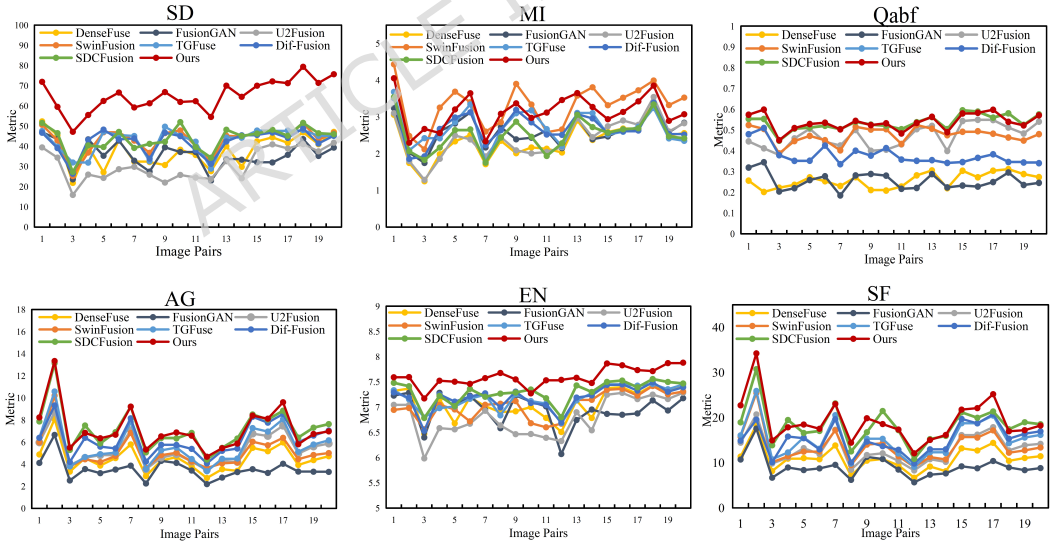


Figure 13. Comparative results of objective metrics for different fusion methods with the RoadScene dataset.

As shown in Fig. 13, this group of experiments is very similar to the situations in Fig. 9. Our method exhibits obvious superiority over others in SD, Qabf, AG, EN, and SF assessment. For MI, our method, SwinFusion, and TGFuse outperform other methods overall. This demonstrates the role of cross-attention again and our method's comprehensiveness.

4.2.3 Computational efficiency comparison

This section further compares the computational efficiency of our method with various SOTA methods, including floating point operations (FLOPs), inference speed (FPS), parameter count (Params), and memory consumption. All comparisons are

conducted under the same hardware environment, with the input image resolution uniformly set to 256*256. The detailed results are shown in Table 1.

Table 1. Computational efficiency comparison of eight fusion methods

Method	FLOPs (G)	FPS	Params (M)	Memory Consumption (MB)
DenseFuse	54.15	105.03	0.07	97.43
FusionGAN	57.09	60.73	1.31	1829
U2Fusion	43.17	65.04	0.66	22709.34
SwinFusion	292.54	3.36	0.97	593.15
TGFuse	840.70	62.03	549.36	901.38
Dif-Fusion	726.10	2.63	434.17	2321.36
SDCFusion	22.99	102.47	0.57	290.37
Ours	259.96	6.37	9.54	677.53

It can be observed from Table 1 that there are significant differences in computational efficiency among different methods. Traditional methods such as DenseFuse, FusionGAN, and U2Fusion have lower FLOPs and higher inference speeds, but their fusion performance is limited. SwinFusion, TGFuse, and Dif-Fusion generally have large computational overheads, making it difficult to meet real-time requirements.

Our method achieves a good balance between computational efficiency and fusion performance. The FLOPs of our method are significantly lower than those of heavy models such as TGFuse and Dif-Fusion. Although its inference speed is not as high as lightweight CNN methods, it is at a leading level among Transformer-based fusion methods. In terms of parameter count, our method is much lower than TGFuse and Dif-Fusion, indicating that the model has high parameter efficiency. The memory consumption is within an acceptable range.

4.3 Ablation experiments

A number of ablation experiments were conducted to verify the significance of our core technologies comprehensively. Twenty pairs of source images from the RoadScene dataset are shown in Fig. 10. The experimental settings are the same as the above contrast experiments. Specifically, the verifications of intra-attention CASTB, STB based feature fusion strategy, and adaptive interaction loss term are reported in Sections 4.3.1, 4.3.2, and 4.3.3, respectively.

4.3.1 The verification and analysis of interaction attention

In our model, the feature extraction stages contain two STBs and one CASTB. The front-end STBs are used to dig intra-attention, and the backend CASTB is responsible for inter-attention. To demonstrate our optimal assembly (denoted as '2+1' below), four contrasted combinations include: 1. two STBs with no CASTB ('2+0'); 2. one STB with one CASTB module ('1+1'); 3. one STB with two CASTB modules ('1+2'); 4. two STBs with two CASTBs modules ('2+2'). The other parts of the network and the loss function remain unchanged. Figure 14 shows average value comparisons of five combinations with six metrics.

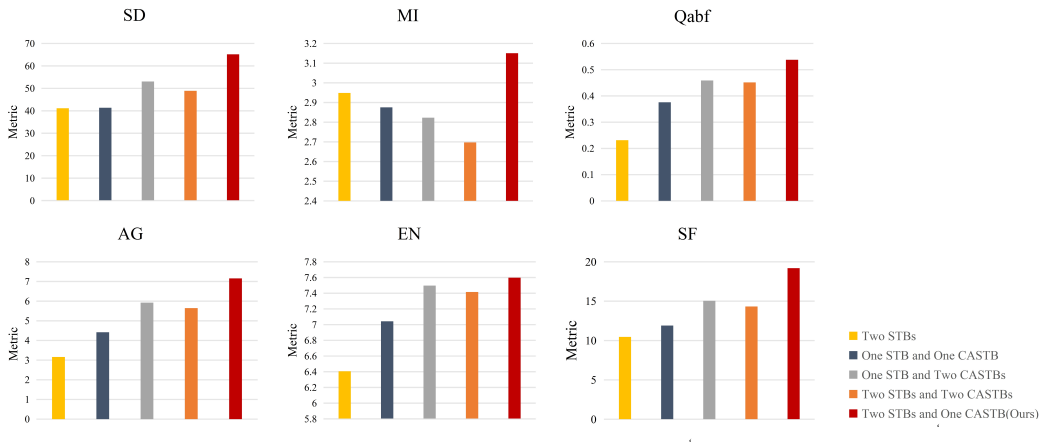


Figure 14. The average value comparisons of five types of feature extraction combinations.

As shown in Fig. 14, the following conclusions can be drawn:

1. In the whole, our ‘1+2’ assembly achieves the best results on all metrics, and ‘2+0’ assembly presents the worst performance on all metrics except MI;

2. Given only one STB is employed, ‘1+2’ is superior to ‘1+1’ assembly generally. This indicates that as the degree of intra-attention increases, the fusion quality can be improved under this situation;

3. When two STBs are employed, it can be seen that a counterintuitive pattern is observed that ‘2+1’ shows a notable advantage over ‘2+2’ assembly in all experiments. It means the number of CASTB is not the more the better in the double-attention framework;

4. Besides the above contrasts, another notable observation is that the rankings on MI show an obvious difference from the other five. Concretely, the ‘2+0’ combination presents better MI assessment than the other four combinations. The reason will be discussed in Section 4.4.

Besides the five combinations shown in Fig. 14, ‘3+1’ and ‘3+2’ combinations were also compared in the experiments. The results demonstrate that our assembly is the best choice. For a more intuitive comparison, two examples of feature heatmaps are shown in Fig. 15.

Figure 15. Two examples of the gradual interaction processes of two modalities.

In each row of Fig. 15, an infrared image patch, its visible image partner, and three heatmaps of infrared features are shown sequentially. These features are the infrared branch of the outputs of CASTB, as shown in Fig. 3(a). Here, we visualize the infrared features to illustrate the gradual interaction process of double-attention. The legend is displayed in the lower right corner. We can note when 0, 1, and 2 CASTBs are used (from (c) to (e)), more and more visible light details are transferred to infrared features, such as the canopy parts in the first row and the headlight in the second line. But on the other hand, the infrared salient objects become blurred gradually, and their contrast degrees decrease gradually, such as the car in the first line and the heat target in the upper left corner in the second line. These examples intuitively illustrate that overly double-attention interactions smooth modality-unique features and lead to the decline of whole salient information. This illustrates that combining the double-attention in a regulatable manner is essential.

Besides the double-attention assemblies contrasted above, other combinations were also tested, such as three STBs and more. Our results indicate that: 1. When more STBs are used, the improvement of performance is not obvious while the running time increases linearly; 2. The number of CASTBs should not exceed half of STBs; this seems to be a good balance point.

4.3.2 The verification of feature fusion strategy

As in Fig. 4, we deploy an STB in the feature fusion stage, rather than using simple concatenation or convolution only. To validate its effectiveness, we design two contrast cases: 1. concatenation of the two source features (Concat.); 2. concatenation followed by convolution (Concat. + Conv.); 3. concatenation followed by a deep convolutional block with the same number of parameters as the STB (Concat.+Conv.-EqualParam);Our strategy is denoted as (Concat. + STB). The contrasted results are presented in Fig. 16.

It is clear from Fig. 16 that our (Concat. + STB) combination achieves the best assessments on all metrics. Although the traditional (Concat. + Conv.) method is better than the simple concatenation,, it is still inferior to the proposed approach. More importantly, the (Concat.+Conv.-EqualParam) baseline, despite having a parameter count comparable to the STB, yields significantly lower performance. This also demonstrates that the performance improvement brought by the STB selected in this paper is not merely due to the increase in the number of parameters. The reason can be explained as further deep integration can help to fuse the extracted features.

Besides the above experiments, another combination (Concat. + 2 STB) was also tested. In this case, the whole luminance of the fusion results decreases notably, and some details are suppressed markedly.

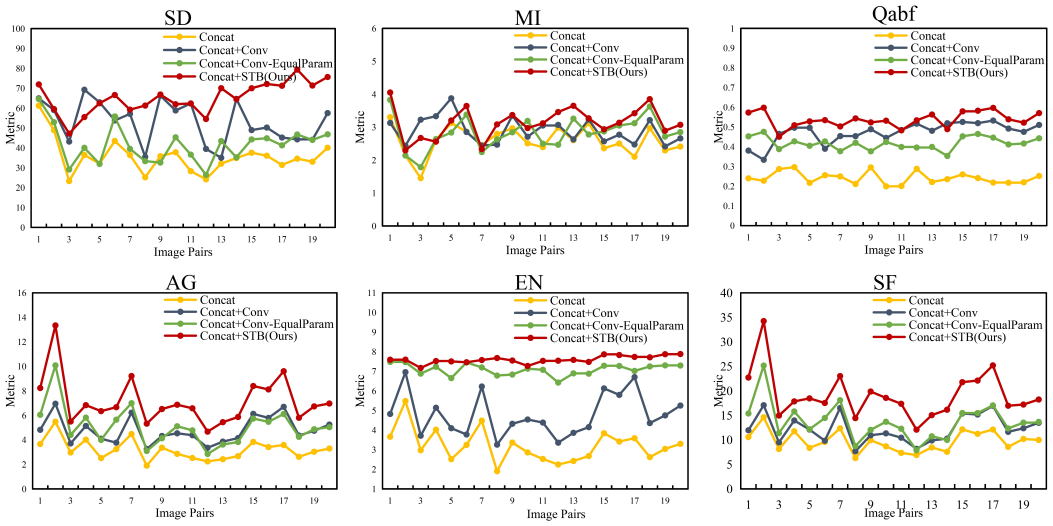


Figure 16. The comparisons of four kinds of feature fusion ways.

4.3.3 The verification of the adaptive interaction loss

As has been validated above, interaction regulation is critical for fusion quality. Here we designed three contrasted loss functions: (1) intensity and gradient losses only; (2) fixed weight SSIM loss and (1); (3) the proposed adaptive loss and (1). The results are presented in Fig. 17.

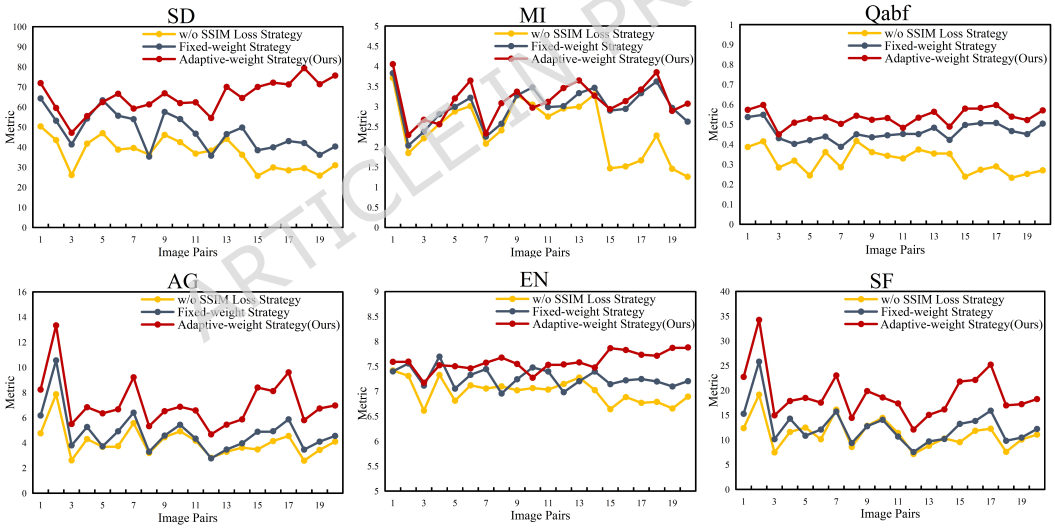


Figure 17. The ablation study of adaptive interaction loss.

As shown in Fig. 17, our loss function achieves the best results in all metrics, followed by the combination (2), and the combination (1) performs the worst. The following conclusions can be drawn:

1. SSIM loss can improve the fusion quality effectively. Combination (2) presents better whole performance than (1) on all metrics;

2. Our combination shows more obvious advantages than combination (2). Compared with the fixed weight SSIM loss, our adaptive loss dynamically adjusts regional weights, guiding the CASTB to regulate interaction degrees, resulting in more retention of infrared targets and visible textures.

4.4 Hyperparameter Sensitivity Analysis

To verify the rationality of the selected loss function weights, this section conducts a sensitivity analysis using the control variable method on the RoadScene dataset. The results are shown in Fig. 18. The selected parameters (10, 10, 15) achieve

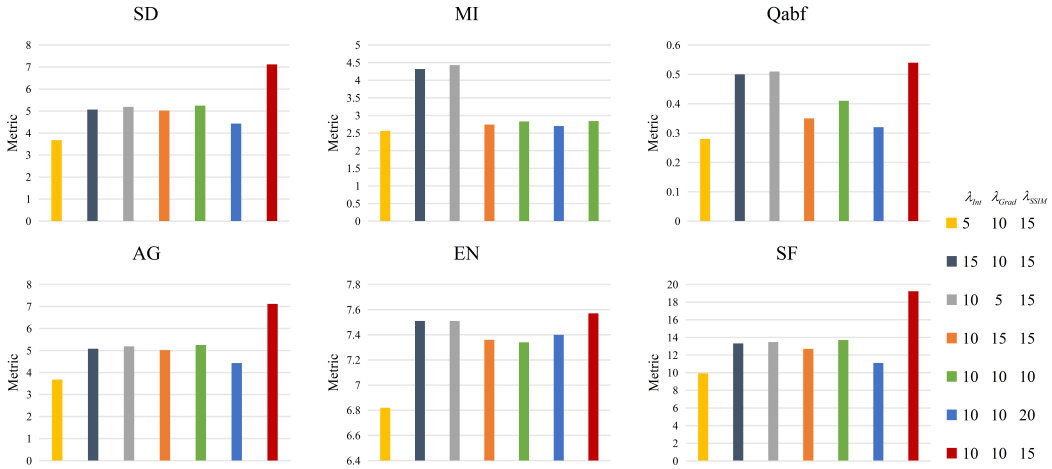


Figure 18. Sensitivity analysis of loss function weights.

the best or suboptimal performance across all metrics, and the model exhibits relatively stable performance under different parameter settings.

4.5 Discussion

In this section, we discuss several related issues based on our experiments.

1. The effects of inter-attention

Our ablation experiments demonstrate fully that incorporating cross-modal inter-attention in the feature extraction stage can improve fusion quality. But on the other hand, the experiments also reveal an important fact that excessive stacking of inter-attention modules will lead to performance degradation, i.e., a performance balance point exists between two kinds of attention.

Specifically, for the IVIF task, cross-modality attention interaction can effectively establish complementary connections between infrared and visible source images. However, it is also a two-edged weapon: an excessive interaction will make the information from the other modality continuously rewrite the features of the current modality. As the illustrations in Fig. 15 show, when two CASTBs are employed, modality-specific salient information begins to be shaded. The success of our model (2 STBs + 1 CASTB) lies in its balanced architecture: the robust self-representations through STB are constructed first, and then proper inter-attention is introduced through CASTB. At the same time, an adaptive interaction loss is crucial for assisting the network training to achieve its ideal targets.

2. The adaptive interaction loss

In the IVIF domain, assigning two source images with equal importance indiscriminately is not optimal due to the fact that in different regions, two sources have divergent saliency. For example, if a flat visible background region contains some infrared details, then the infrared details would be suppressed due to the equal loss weights. In the double-attention fusion framework, this drawback is magnified due to the deep interaction between two heterogeneous modalities. In order to interact with two modalities in a controllable way, we design the interaction loss term to assist the network model. From the results of the contrast and ablation experiments, our method achieves the expected goal.

3. The objective assessment metrics

Assessing the fusion quality comprehensively and purposefully is a complex issue because of the ambiguity of ‘saliency information’ and the diversity of evaluation viewpoints. Thus, joint assessment with multiple metrics is widely used. From our experiments, two conclusions can be seen:

(1) As shown in Fig. 9, Fig. 13, our method shows significant advantages on SD, Qabf, AG, EN, and SF in all experiments. Because these metrics measure the changes in the fusion image essentially, this manifests that our results contain more edges and texture details.

(2) In regard to MI, the superiority of our method is not clear, as in other metrics, similar situations also occur in ablation experiments. This is not indicative of inferior fusion quality but rather reflects a fundamental trade off. MI measures the overall statistical dependency between the source images and the fused image. Driven by the factors of employing CASTB in the feature extraction stage, STB in the fusion stage, and adaptive interaction loss in the training stage jointly, more salient information is mined and integrated, thus the overall distribution similarity based MI is affected to some extent.

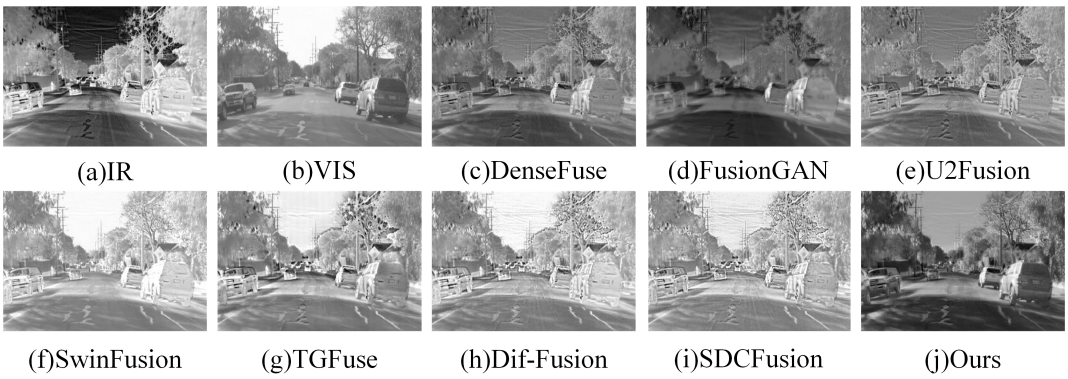


Figure 19. A representative case comparing MI values and visual quality.

To further prove this point, a representative case is shown in Fig. 19. In this case, the MI value of our method is 2.04. This is lower than FusionGAN (2.17), SwinFusion (2.10), Dif-Fusion (2.23), and SDCFusion (2.54). However, our fused image shows clearer thermal targets and richer texture details. Some comparative methods achieve higher MI values. But their fusion results fail to preserve salient structures effectively. Severe over-smoothing or information loss occurs. Our method moderately sacrifices MI values. The goal is to highlight local salient features. This trade-off is necessary for achieving superior visual quality.

In the whole, our method successfully achieves a compromise between overall similarity retention and local saliency highlighting, significantly advancing detail preservation and contrast while maintaining a robust level of overall structural fidelity.

5 Conclusion

To exploit the attention mechanism for the cross-modal IVIF fusion task, this paper proposes a novel IVIF method based on a double-attention mechanism and adaptive interaction loss. Our novelties include a more reasonable double-attention fusion framework and solving cross-modal interaction regulation effectively. Concretely, we devise a two-step cross-modal double-attention strategy for more complete complementary feature extraction; further, to take full advantage of these complementary features, an STB is utilized in the feature fusion stage. This collaborative architecture ensures our fusion results retain more salient content. Moreover, an adaptive interaction loss was devised to regulate the interaction degree between two kinds of attentions more accurately. By adjusting the loss weights of two source images according to their regional salient degrees adaptively, the optimal double-attention interaction is achieved, and more saliency information is retained. In the experiments, we compared our method with seven SOTA methods on two datasets. A large number of subjective and objective comparisons demonstrate the superiority of the proposed method. A series of ablation experiments was carried out to verify the effectiveness of the designed modules. Furthermore, several related issues are also discussed in depth.

Our future work includes designing more effective interaction control methods and exploiting them in other multi-modal vision tasks. In addition, the investigation of dedicated fusion methods for low-quality source images is also a valuable direction for future research.

References

1. Zhang, H., Xu, H., Tian, X., Jiang, J. & Ma, J., Image fusion meets deep learning: A survey and perspective, *Information Fusion* 76 (2021) 323–336.
2. Karim, S. et al. ReMamba: a hybrid CNN-Mamba aggregation network for visible-infrared person re-identification. Sci Rep 14, 29362 (2024).
3. Vivone, G., Multispectral and hyperspectral image fusion in remote sensing: A survey, *Information Fusion* 89, 405–417 (2023).
4. Li, J. et al. AttFeat: Attention-based features for infrared and visible remote sensing image matching, *IEEE Geoscience and Remote Sensing Letters* 22, 1–5 (2025).
5. Voronin, V. et al. Deep visible and thermal image fusion for enhancement visibility for surveillance application, in: *Security + Defence* (2022).

6. Yadav,G. & Yadav, D., Contrast enhancement of region of interest of backlit image for surveillance systems based on multi-illumination fusion, *Image and Vision Computing* 135, 104693 (2023).
7. Yuan, Q. et al. Enhanced target tracking algorithm for autonomous driving based on visible and infrared image fusion, *Journal of Intelligent and Connected Vehicles* 6, 237–249 (2023).
8. Wang, D., Liu, J., Liu, R. & Fan, X., An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection, *Information Fusion* 98,101828 (2023).
9. Bineeshia, J. & Kumar, B.V. AIR-GANet: multi-head attention integrated residual dense block based generative adversarial network for visible and infrared image fusion. *Sci Rep* 15, 39464 (2025).
10. Zhuang, C. et al. PHFuse: Unsupervised color visible and infrared image fusion with preserved hue. *Sci Rep* 15, 31458 (2025).
11. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale, Preprint at https://arxiv.org/abs/2010.11929 (2021).
12. Ahmad, I. et al. A multiscale transformer with spatial attention for hyperspectral image classification. *Sci Rep* (2026).
13. Ma, J. et al. SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer, *IEEE/CAA Journal of Automatica Sinica* 9,1200–1217 (2022) .
14. Rao, D. et al. TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network, *IEEE Transactions on Image Processing* (2023).
15. Karacan, L., Multi-image transformer for multi-focus image fusion, *Signal Processing: Image Communication* 119,117058 (2023).
16. Wang, Z. et al. SwinFuse: A residual Swin transformer fusion network for infrared and visible images, *IEEE Transactions on Instrumentation and Measurement* 71,117058 (2022).
17. Chen, X., Xu, S., Hu, S. & Ma, X., MGFA: A multi-scale global feature autoencoder to fuse infrared and visible images, *Signal Processing: Image Communication* 128,117168 (2024)
18. Li, H. et al. DenseFuse: A fusion approach to infrared and visible images, *IEEE Transactions on Image Processing* 28, 2614–2623 (2019).
19. Li, H., Wu, X. & Kittler, J., RFN-Nest: An end-to-end residual fusion network for infrared and visible images, *Information Fusion* 73, 72–86 (2021).
20. Zhang, Z., Wu, X. & Xu, T., FPNFuse: A lightweight feature pyramid network for infrared and visible image fusion, *IET Image Processing* 16, 2308–2320 (2022).
21. Wang, H., Lu, X., Wu, Z., Li, R. & Wang, J., Infrared and visible image fusion based on autoencoder network, *IET Image Process* 19, 70086 (2025).
22. Liu, R. et al. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion, *IEEE Transactions on Image Processing* 30, 1261–1274 (2021).
23. Liu, Y., Chen, X., Cheng, J., Peng, H. & Wang, Z., Infrared and visible image fusion with convolutional neural networks, *International Journal of Wavelets, Multiresolution and Information Processing* 16, 1850018 (2018).
24. Xu, H. et al. U2Fusion: A unified unsupervised image fusion network, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 502–518 (2022).
25. Long, Y., Jia, H., Zhong, Y., Jiang, Y. & Jia, Y., RXDNFuse: A aggregated residual dense network for infrared and visible image fusion, *Information Fusion* 69, 128–141 (2021).
26. Li, H. et al. 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, 4070–4083 (2021).
27. Mustafa, H., Yang, J., Mustafa, H. & Zareapoor, M., Infrared and visible image fusion based on dilated residual attention network, *Optik* 224 , 165409 (2020).
28. Luo, Y., He, K., Xu, D., Yin, W. & Liu, W., Infrared and visible image fusion based on visibility enhancement and hybrid multiscale decomposition, *Optik* 258, 168914 (2022).
29. Xu, J., Liu, Z. & Fang, M., An infrared and visible image fusion network based on multi-scale feature cascades and non-local attention, *IET Image Process* 18, 2114–2125 (2024).

30. Hu, X., Liu, Y., Yang, F., PFCFuse: a Poolformer and CNN fusion network for infrared-visible image fusion, *IEEE Transactions on Instrumentation and Measurement* 73, 1-14 (2024).

31. Ma, J. et al. FusionGAN: A generative adversarial network for infrared and visible image fusion, *Information Fusion* 48, 11–26 (2019).

32. Zarimeidani, M., Amirabadi, A., Amiri, N., Ahanian, I., & Es'haghi, S., Infrared and visible image fusion using GAN with fuzzy logic and Harris Hawks optimization. *Sci Rep* 16, 70 (2026).

33. Lebedev, M., Komarov, D., Vygolov & Vizilter, Y., Multisensor image fusion based on generative adversarial networks, in: *Image and Signal Processing for Remote Sensing XXV*, 111551T (2019).

34. Zhang, D., Yong, D., Zhao, J., Zhou, Z. & Yao, R, Structural similarity preserving GAN for infrared and visible image fusion, *Multiresolution and Information Processing* 19, 2050063 (2020) .

35. Ma, J. et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion, *IEEE Transactions on Instrumentation and Measurement* 70, 1–14 (2021).

36. Ma, J. et al. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, *IEEE Transactions on Image Processing* 29, 4980–4995 (2020).

37. Shen, S. et al. ADF-Net: Attention-guided deep feature decomposition network for infrared and visible image fusion, *IET Image Process.* 18, 2774–2787 (2024).

38. Tang, W. et al. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer, *IEEE Transactions on Multimedia* 25, 5413–5428 (2023).

39. Li, J. et al. CGTF: Convolution-guided transformer for infrared and visible image fusion, *IEEE Transactions on Instrumentation and Measurement* 71, 1–14 (2022).

40. Yang, X. et al. DGLT-Fusion: A decoupled global–local infrared and visible image fusion transformer, *Infrared Physics & Technology* 128, 104522 (2023).

41. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows, Preprint at https://arxiv.org/abs/2103.14030 (2021).

42. Golkhatmi, B., Houshmand, M. & Hosseini, S., A multi-scale attention-based Swin transformer model for medical images segmentation. *Sci Rep* 15, 38893 (2025).

43. Ke, A., Luo, J. & Cai, B., UNet-like network fused swin transformer and CNN for semantic image synthesis. *Sci Rep* 14, 16761 (2024).

44. Yang, R., Liu, K. & Liang, Y., A fusion-attention swin transformer for cardiac MRI image segmentation, *IET Image Processing* 18, 105–115 (2024).

45. Li, H. et al. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion, *IEEE Transactions on Pattern Analysis and Machine Intelligenc* 47,3673-3690 (2025).

46. Liu, G., Qiu, J. & Yuan, Y. A Multi-Level SAR-Guided Contextual Attention Network for Satellite Images Cloud Removal, *Remote Sensing* 16, 4767 (2024).

47. Wang, Z. et al. Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing* 13, 600–612 (2004).

48. Ma, K., Zeng, K. & Wang, Z. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.* 24, 3345-3356 (2015).

49. Tang, L., Yuan, J. & Ma, J., Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, *Information Fusion* 82, 28–42 (2022).

50. Yue, J. et al. Dif-Fusion: Toward High Color Fidelity in Infrared and Visible Image Fusion With Diffusion Models, *IEEE Transactions on Image Processing* 32, 5705-5720 (2023).

51. Liu, X., Huo, H., Li, J., Pang, S. & Zheng, B., A semantic-driven coupled network for infrared and visible image fusion, *Information Fusion* 108,1556-2535 (2024).

52. Toet, A., The TNO multiband image data collection, *Data Brief* 15, 249–251 (2017).

53. Tian, Y., Carballo, A., Li, R. & Takeda, K., Road scene graph: A semantic graph-based scene representation dataset for intelligent vehicles, Preprint at https://arxiv.org/abs/2011.13588(2020).

54. Qu, G. et al. Information measure for performance of image fusion, *Electronics Letters* 38, 313–315 (2002) .

55. Han, Y., Cai, Y., Cao, Y. & Xu, X., A new image fusion performance metric based on visual information fidelity, *Information Fusion* 14, 127–135 (2013).

56. Liu, Z. et al. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study, *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 94–109 (2012). MATLAB Code at https://github.com/zhengliu6699/imageFusionMetrics.

Acknowledgements

We would like to thank Professor Liu Zheng for providing fusion quality objective assessment toolbox.

Funding

This work was supported by National Natural Science Foundation of China under Project Number 61274021 and 61902282.

Author contributions statement

Z.W. designed and implemented the fusion framework, conducted experiments, and performed data curation. Y.H. and B.Z. conceived the research idea and supervised the project. Z.W. and Y.H. wrote the manuscript. All authors reviewed, edited, and approved the final manuscript.

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Figure legends

Fig. 1 Windrose map comparisons of the proposed method and seven SOTA methods on two IVIF datasets. The six evaluation metrics include standard deviation (SD), mutual information (MI), Qabf, spatial frequency (SF), entropy (EN) and average gradient (AG).

Fig. 2 Illustration of the shifted window approach in ST. This diagram demonstrates the Shifted Window based Multi-head Self-Attention (SW-MSA) mechanism. The left part (Layer l) shows regular non-overlapping window partitioning, and the right part (Layer l+1) shows the windows after shifting by (M/2, M/2) pixels, which enables cross-window connections and expands the attention receptive field.

Fig. 3 Overall network architecture of the proposed model. (a) The complete three-stage pipeline: Dual-Attention Feature Extraction, Feature Fusion, and Reconstruction. (b) Details of the Swin-Transformer Layer (STL) for intra-attention. (c) Details of the Cross-attention Swin-Transformer Layer (CASTL) for inter-attention interaction.

Fig. 4 Comparison of three feature fusion strategies. (a) Simple concatenation (Concat.). (b) Concatenation followed by a convolutional layer (Concat.+Conv.). (c) The proposed strategy: concatenation followed by a Swin-Transformer Block (Concat.+STB).

Fig. 5 Illustration of the cross-modal shifted window attention. Similar to Fig.2, it is applied to cross-modal interaction, showing how Cross-Modality W-MSA (left) and SW-MSA (right) compute attention between infrared (IR) and visible (VIS) modal features across regular and shifted windows, thus facilitating cross-window boundary feature interaction.

Fig. 6 Twenty pairs of source images selected from the TNO dataset. These image pairs with diverse scenes containing salient thermal targets and rich textures are used for the testing and comparison in Section 4.2.1.

Fig. 7 Fusion results of the ‘Kaptein_1654’ image pair. Subjective visual comparison of fused images generated by seven SOTA methods and the proposed method. Key areas are marked with boxes to facilitate the visual assessment of detail preservation, target saliency and noise level.

Fig. 8 Fusion results of the ‘barbed_wire’ image pair. Another example of subjective comparison. The fused images are evaluated based on clarity, thermal target contrast and overall naturalness, and the proposed result achieves balanced integration of textures and targets.

Fig. 9 Comparative results of objective metrics for different fusion methods on the TNO dataset. Line charts displaying the scores of six metrics (SD, MI, Qabf, AG, EN, SF) across 20 test image pairs (Fig.6). Each subplot corresponds to one metric, and the curves of different methods are distinguished by colors and markers.

Fig. 10 Twenty pairs of source images selected from the RoadScene dataset. These image pairs featuring traffic scenes with moving vehicles and complex backgrounds are used for the testing in Section 4.2.2 and ablation studies.

Fig. 11 Fusion results of the 'FLIR_04514' image pair. Subjective evaluation focusing on car targets. Comparisons highlight the issues of other methods such as blurriness, artifacts or loss of distant mountain details, while the proposed method maintains clear targets and rich contextual information.

Fig. 12 Fusion results of the 'FLIR_06832' image pair. This example contains multiple objects, and the comparison demonstrates the proposed method's capability to preserve the contours and details of all salient objects against complex backgrounds.

Fig. 13 Comparative results of objective metrics for different fusion methods on the RoadScene dataset. Similar to Fig.9, this set of line charts presents the metric scores for 20 RoadScene image pairs (Fig.10), and the results further verify the advantages of the proposed method, especially in detail- and contrast-related metrics.

Fig. 14 Average value comparisons of five types of feature extraction combinations. A bar chart showing the average scores of six metrics for five different configurations of STB and CASTB modules (e.g., '2+0', '1+1', '2+1'), which quantitatively validates that the proposed '2+1' configuration achieves the best overall performance.

Fig. 15 Two examples of the gradual interaction processes of two modalities. Each row includes: (a) an infrared patch, (b) its corresponding visible patch, and heatmaps of infrared features after interacting with (c) 0, (d) 1, and (e) 2 CASTB modules.

Fig. 16 Comparisons of four feature fusion methods. A line chart comparing the average metric performance of three fusion strategies in Fig.4: Simple Concat., Concat.+Conv., and the proposed Concat.+STB. The results demonstrate the superiority of the STB-based fusion strategy.

Fig. 17 Ablation study of adaptive interaction loss. A line chart comparing the performance under three loss configurations: (1) Intensity+Gradient loss only, (2) Fixed-weight SSIM loss added, (3) The proposed adaptive SSIM loss added. This validates the effectiveness of the proposed adaptive loss in improving fusion quality.

Fig. 18 Sensitivity analysis of loss function weights. A line chart shows the performance trends of six evaluation metrics (SD, MI, Qabf, AG, EN, SF) under different weight combinations. Using the control variable method, each subplot illustrates the impact of a single weight parameter (λ_{Int}, λ_{Grad}, or λ_{SSIM}) while fixing the other two.

Fig. 19 A representative case comparing MI values and visual quality. Visual comparisons of fusion results from different methods on a representative image pair.