Chinese Journal of Information Fusion
http://dx.doi.org/10.62762/CJIF.2025.822280

RESEARCH ARTICLE

Self-supervised Segmentation Feature Alignment for Infrared and Visible Image Fusion

Weitao Qiu[1], **Wenda Zhao**[1,*] **and Haipeng Wang**[2]

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[2] Unit 92728 of PLA, Shanghai 200436, China

Abstract

Existing deep learning-based methods for infrared and visible image fusion typically operate independently of other high-level vision tasks, overlooking the potential benefits these tasks could offer. For instance, semantic features from image segmentation could enrich the fusion results by providing detailed target information. However, segmentation focuses on target-level semantic feature information (e.g., object categories), while fusion focuses more on pixel-level detail feature information (e.g., local textures), creating a feature representation gap. To address this challenge, we propose a self-supervised segmentation feature alignment fusion network (SegFANet), which aligns target-level semantic features from segmentation tasks with pixel-level fusion features through self-supervised learning, thereby bridging the feature gap between the two tasks and improving the quality of image fusion. Extensive experiments on the WHU and Potsdam datasets show our method's effectiveness, outperforming the state-of-the-art methods.

Keywords: image fusion, self-supervised segmentation feature alignment, feature interaction, deep learning.

1 Introduction

Infrared images, formed by capturing thermal radiation, offer strong anti-interference but suffer from low resolution and lack fine details. In contrast, visible images, formed by utilizing light reflected from objects, exhibit high spatial resolution and provide abundant texture details and color information. However, their performance is significantly degraded under low-light or other extremely harsh conditions, thereby compromising target saliency [6, 7]. Consequently, infrared and visible images exhibit inherent complementarity. By fusing these two modalities, the resulting composite image preserves the abundant textural details from the visible image while simultaneously highlighting the salient target information captured by the infrared image. Infrared and visible image fusion technology has demonstrated extensive applicability across diverse domains including military reconnaissance [1], security surveillance [2], video surveillance [3], person re-identification [4] and remote sensing [5].

Academic Editor:
Fengbao Yang

Submitted: 26 May 2025
Accepted: 20 August 2025
Published: 21 September 2025

Vol. 2, **No.** 3, 2025.
10.62762/CJIF.2025.822280

***Corresponding author:**
✉ Wenda Zhao
zhaowenda@dlut.edu.cn

Citation
Qiu, W., Zhao, W., & Wang, H. (2025). Self-supervised Segmentation Feature Alignment for Infrared and Visible Image Fusion. *Chinese Journal of Information Fusion*, 2(3), 223–236.

Image fusion focuses on pixel-level detail information, but rarely integrates target-level semantic information. In contrast, image segmentation focuses on target-level semantic information such as object categories. Image segmentation can provide target-level semantic information for image fusion, helping it better fuse the target area during the fusion process. Therefore, the core task of this study is to utilize leverage the advantages of multi-task learning, using the segmentation task to provide target-level semantic information for the fusion task, thereby guiding the fusion process to preserve and enhance target regions and improve the quality of image fusion. To this end, we focus on solving the problem of "how segmentation task can assist fusion task" and bridging the feature gap between the two tasks.

Existing deep learning-based image fusion methods can be roughly divided into the following four categories: CNN-based methods [19, 22–25, 36–38], AE-based methods [11, 13, 14, 20, 32], GAN-based methods [8, 15, 16, 21, 33–35] and methods that jointly learn image fusion and high-level vision tasks [9, 17, 18, 39]. The core idea of CNN-based methods is to design network structures and loss functions so that the model can automatically learn the optimal fusion strategy and achieve end-to-end feature extraction, feature fusion, and feature reconstruction. The core idea of AE-based methods is to achieve feature extraction and image reconstruction by training an autoencoder network. GAN-based methods generate high-quality fused images through adversarial learning between the generator and the discriminator. In addition, some studies attempt to jointly optimize image fusion and high-level vision tasks by designing a loss function based on multi-task learning, but it is still difficult to overcome the fundamental problem of the feature gap caused by hierarchical differences. To address this problem, we propose a self-supervised segmentation feature alignment fusion network for infrared and visible image fusion (SegFANet). This network uses a self-supervised approach to achieve segmentation feature alignment. By converting the semantic-level features of segmentation into pixel-level features suitable for image fusion, it bridges the feature gap between the two tasks and achieves collaborative collaboration.

Specifically, we design an image reconstruction module whose core function is to process the semantic-level features output by the segmentation network. This module is trained using a self-supervised strategy and can reconstruct the semantic-level features generated by image segmentation into pixel-level features suitable for image fusion. In detail, this module converts a semantic-level feature into a reconstructed image through convolution, then uses the original image as

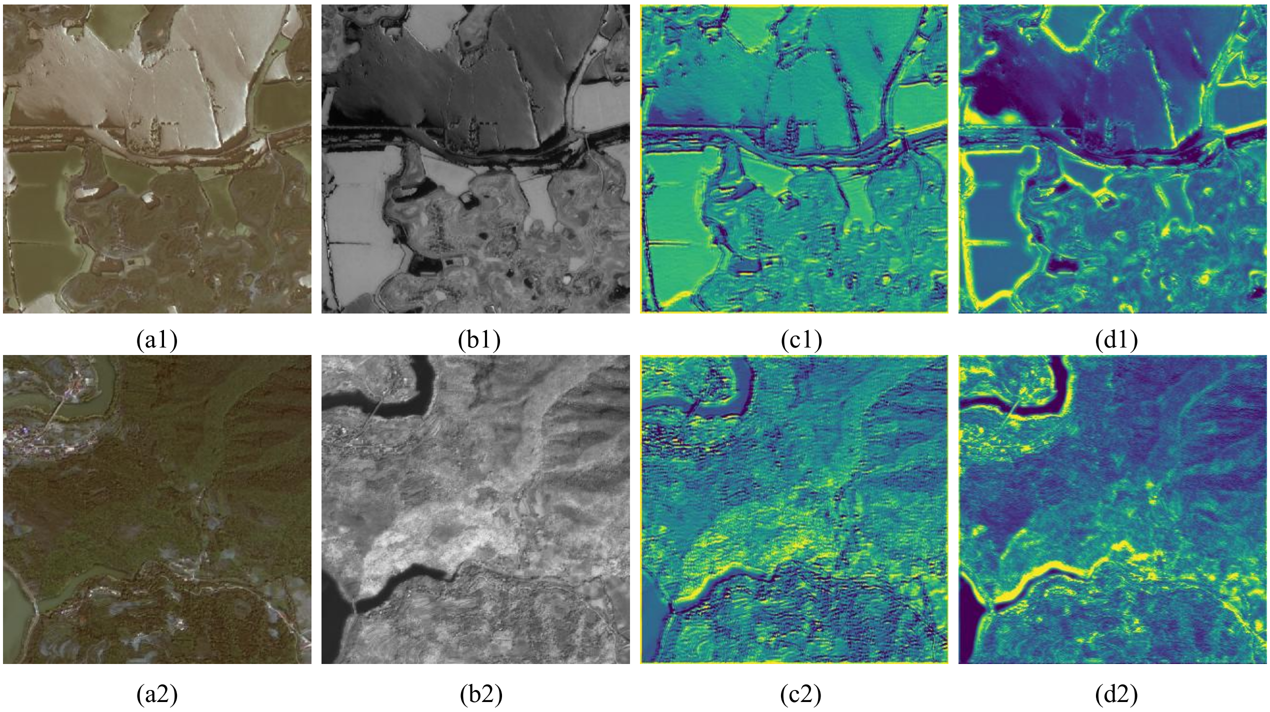


Figure 1. Visualizations of feature distributions before and after alignment. (a1-d1) and (a2-d2) represent visible image, infrared image, the visualization result before alignment, and the visualization result after alignment, respectively.

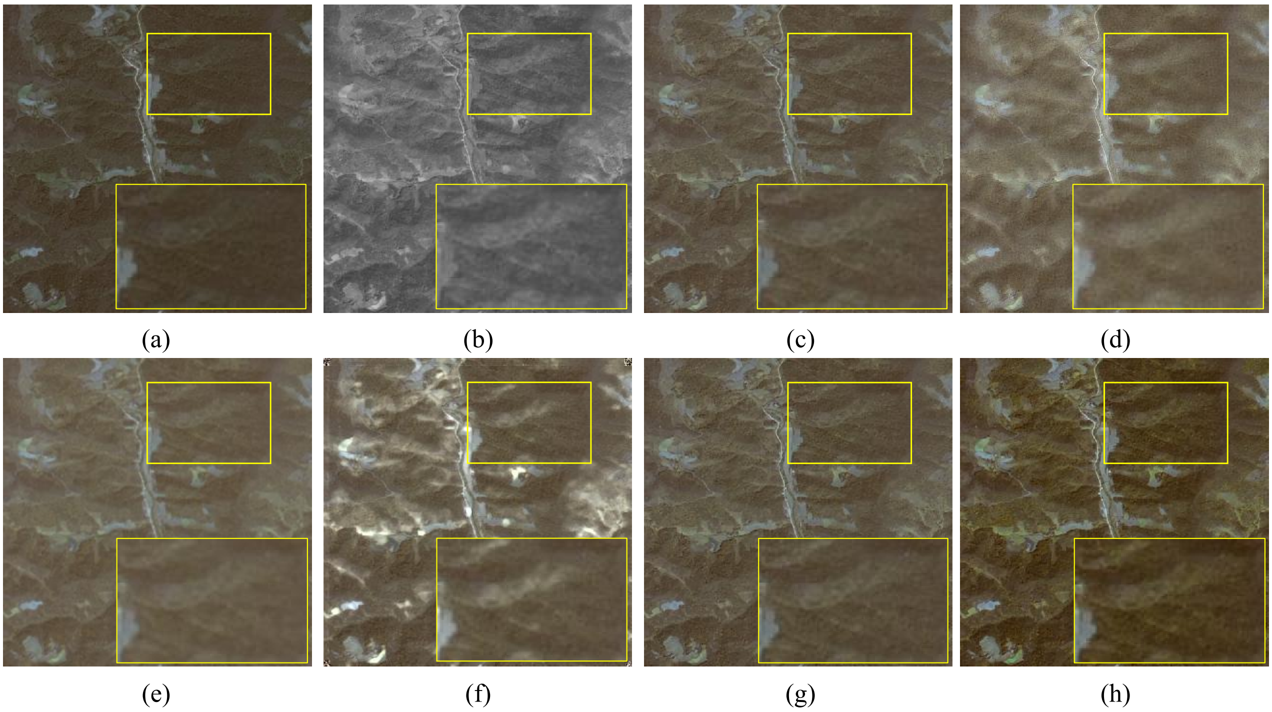


Figure 2. Visual comparison with the state-of-the-art methods. (a-h) are visible image, infrared image, UMFusion, LiMFusion, ITFuse, Tardal, YDTR and our model.

a reference label to constrain through reconstruction loss. This process bridges the feature gap between segmentation and fusion, enabling effective feature alignment. As shown in Figure 1, the feature map before alignment only retains category-based semantic information, with blurred edges and missing pixel-level details; after alignment, the feature map exhibits pixel-level details. Building on this module, we incorporate an attention mechanism to facilitate feature interaction between the two tasks, enabling effective collaboration and complementary enhancement between image segmentation and fusion processes, thereby improving the quality of image fusion. As illustrated in the locally enlarged areas of Figure 2, our method not only preserves the texture details and color information from visible images but also successfully integrates the thermal radiation information from infrared images. The main contributions are summarized as follows:

- We design an image reconstruction module that bridges the feature gap between image fusion and segmentation tasks by converting semantic-level features from the segmentation network into pixel-aligned feature representations suitable for image fusion.

- We introduce an attention mechanism to promote feature interaction between the two tasks. In this way, the segmentation task can provide semantic information for the fusion task, better improving the performance of the fusion network and generating high-quality fused images.

- The experimental results demonstrate that our method has certain effectiveness in performance. As shown in Figure 2, compared with the state-of-the-art methods, our fusion results demonstrate superior performance.

2 Related Work

2.1 CNN-based fusion methods

CNN-based fusion methods can automatically learn the features of the input image by designing a specific network structure and loss function, and fuse these features to generate a high-quality fused image. This process is mainly divided into three steps: feature extraction, feature fusion and image reconstruction. For example, Zhang et al. [22] propose IFCNN, which first uses two convolutional layers to extract salient features from multiple input images, then selects appropriate fusion rules to fuse the extracted features, and reconstructs the fused image through two convolutional layers. Ma et al. [23] propose STDFusionNet, which uses salient object masks to assist fusion tasks. Considering illumination, Tang et al. [24] propose a progressive image fusion network based on illumination perception. Wang et al. [25] propose UMFusion, which generates pseudo-infrared images through a crossmodality perceptual style transfer network (CPSTN) and uses a multi-level refinement registration network (MRRN) for image registration. Finally, the feature interaction fusion module (IFM) is used to adaptively select features for fusion in the dual-path interaction fusion network (DIFN). Furthermore, transformer has shown excellent performance in the visual field due to its powerful modeling ability. Therefore, Tang et al. [19] propose YDTR, which obtains local features and important contextual information through the Y-shaped dynamic transformer module.

2.2 AE-based fusion methods

AE-based fusion methods achieve feature extraction and image reconstruction by using pre-trained autoencoders, and use manually designed fusion rules in the fusion process. Li et al. [13] propose DenseFuse. Unlike traditional convolutional networks, the encoder of DenseFuse consists of convolutional layers, fusion layers and dense blocks. Li et al. [14] propose NestFuse, which introduces a nest connection architecture and retains multi-scale feature information. Li et al. [20] propose an end-to-end fusion network architecture (RFN-Nest), using RFN to replace traditional methods.

2.3 GAN-based fusion methods

The Generative Adversarial Network (GAN) consists of a generator and a discriminator. GAN-based fusion methods use the adversarial training mechanism of the generator and the discriminator to extract features from the input image and generate the fused image. For example, Ma et al. [15] propose FusionGAN, which constructs an adversarial game mechanism between the generator and the discriminator. In addition, Ma et al. [16] propose a dual-discriminator conditional generative adversarial network (DDcGAN), which achieves the fusion of infrared and visible images with different resolutions through adversarial training between generators and two discriminators.

However, most of these existing deep learning-based image fusion methods are independent of other high-level visual tasks, such as object detection [30] and image segmentation [31]. Currently, Tang et al. [18] propose SeAFusion, which cascades the image fusion module and the semantic segmentation module, and designs a loss based on multi-task learning to constrain the fusion network. But existing multi-task learning methods are mainly applicable to tasks at the same level. As two vision tasks, image fusion and image segmentation have significant differences in feature representation, so bridging the feature gap between the two tasks is still a difficult problem. To bridge this gap, our method adopts the self-supervision idea to convert the segmentation features into pixel-level features that match the image fusion task, thereby narrowing the difference in feature representation between the two tasks.

3 Methodology

3.1 Overview

Our SegFANet framework is shown in Figure 3, which consists of three sub-networks: the segmentation network aims to extract target-level features, while the fusion network focuses on pixel-level feature extraction and integration. In order to take advantage of the complementary advantages of multi-task learning, we introduce stage-interactive networks (SINets) between the encoder stages of the two networks. The stage-interactive network aims to assist the fusion network with the semantic information of the segmentation network to help the fusion network better understand the image content. This is achieved through 3 key modules: image reconstruction module (IRM), cross attention module (CAM) and feature fusion module (FFM).

Specifically, SegFANet conducts cross-task feature interactions between the corresponding encoder levels of the segmentation network and the fusion network through n stage-interactive networks. In each network, first, segmentation features and fusion features are extracted from the infrared and visible inputs:

$$f_{e,i}^{S} = E_i^{S}(I_{\text{ir}}, I_{\text{vis}}), \tag{1}$$

$$f_{e,i}^{F} = E_i^{F}(I_{\text{ir}}, I_{\text{vis}}), \tag{2}$$

where I_{ir} and I_{vis} denote the infrared and visible input images, $E_i^S(\cdot)$ and $E_i^F(\cdot)$ are the i-th encoders of the segmentation and fusion branches, $f_{e,i}^S$ and $f_{e,i}^F$ represent their corresponding encoder features.

Then, we construct IRM based on a self-supervision mechanism, which transforms the target-level semantic features $f_{e,i}^S$ from the segmentation network into pixel-level features f_i^R to bridge the feature gap between image fusion and segmentation:

$$f_i^R = R_i(f_{e,i}^S), \tag{3}$$

where $R_i(\cdot)$ denotes the function of the i-th IRM, which includes four "Convolution with 3×3 kernel +

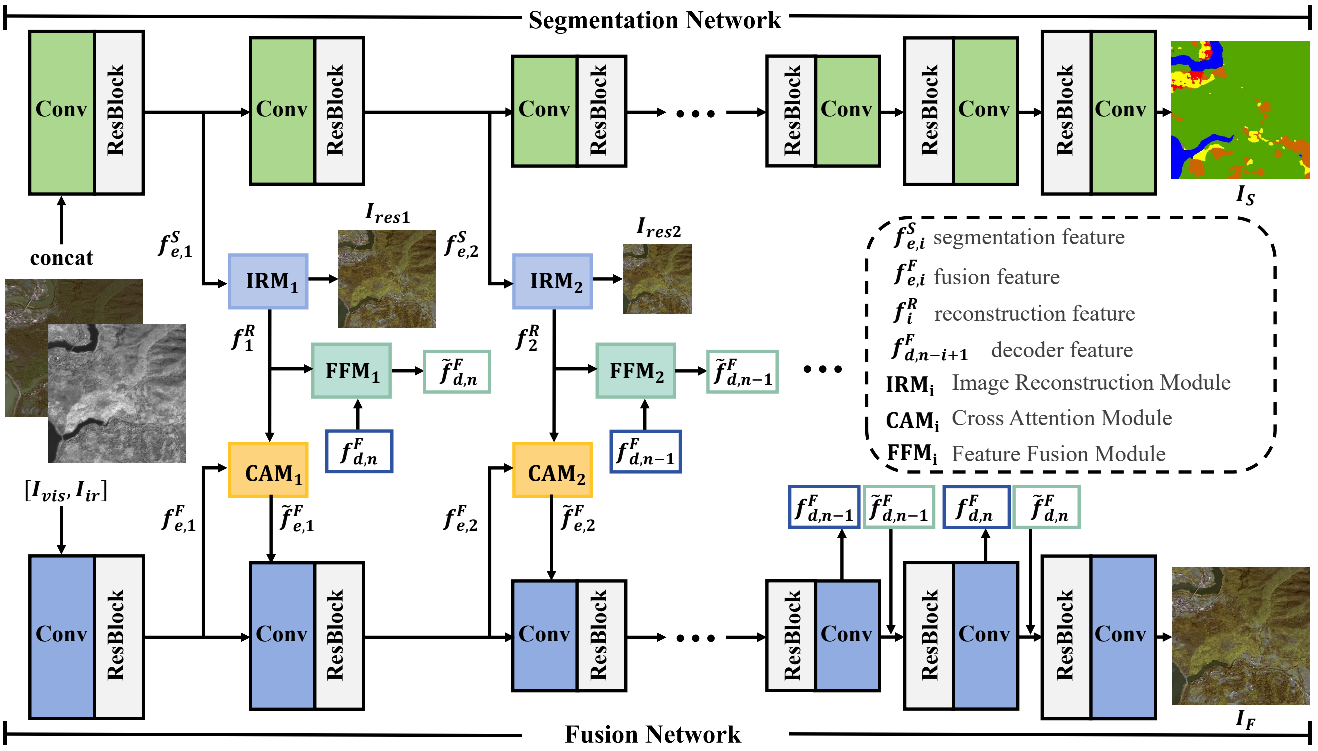


Figure 3. The overall workflow of the proposed model. The upper part describes the architecture of the segmentation network. The middle part details the proposed stage-interactive network. The lower part outlines the structure of the fusion network.

ReLU" layers, and $f^S_{e,i}$ denotes the segmentation feature produced by the segmentation branch's i-th encoder and serves as the input for IRM.

In addition, CAM takes the features $f^F_{e,i}$ and f^R_i obtained by the fusion network and IRM as inputs, and interacts to obtain a new fusion feature $\tilde{f}^F_{e,i}$, which can be formulated as:

$$\tilde{f}^F_{e,i} = S_i(f^F_{e,i}, f^R_i), \tag{4}$$

where $S_i(\cdot)$ denotes the function of the i-th CAM, $f^F_{e,i}$ is the fusion feature, and f^R_i is the reconstructed feature.

Moreover, we develop the FFM, which takes the feature f^R_i and the feature $f^F_{d,n-i+1}$ from the fusion decoder stage at the corresponding resolution as inputs and performs feature fusion. This process enhances the fusion network decoder's ability to understand semantic information, thereby enabling the fusion network to generate high-quality and semantically rich fusion images, which can be formulated as:

$$\tilde{f}^F_{d,n-i+1} = F_i(f^R_i, f^F_{d,n-i+1}), \tag{5}$$

where $F_i(\cdot)$ denotes the function of the i-th FFM. In the following subsections, we introduce the detailed architectures of CAM and FFM, respectively. In addition, we elaborate on the design of the loss function, which plays a crucial role in guiding the model to improve image fusion quality.

3.2 Cross Attention Module

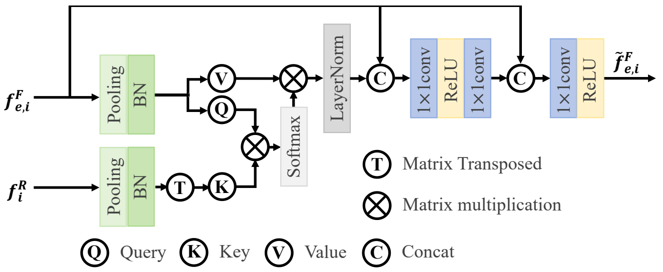


Figure 4. The structure of the cross attention module.

The specific structure of the cross attention module (CAM) is shown in Figure 4. In this module, the input feature $f^F_{e,i}$ is first processed by adaptive average pooling to compress the spatial dimension, and then the Query Q and Value V are generated through the batch normalization layer. At the same time, the same adaptive average pooling and batch normalization operations are performed on another input feature f^R_i to generate a Key K, which provides a clear semantic prior. By calculating the similarity between K and Q, the fused feature is guided to focus on the target area.

Then the softmax function is applied for normalization to generate the attention weight. Finally, the attention weight is used to compute a weighted sum of the V, completing the feature fusion operation. The interacted feature is concatenated with the original fused feature $f_{e,i}^F$ to obtain the final fused feature $\tilde{f}_{e,i}^F$. The concatenated feature $\tilde{f}_{e,i}^F$ is then directly fed into the convolutional layer of the next stage in the image fusion encoder as its input. This process can be defined as:

$$Q = \text{BN}(\text{AAP}(f_{e,i}^F)), \tag{6}$$

$$V = \text{BN}(\text{AAP}(f_{e,i}^F)), \tag{7}$$

$$K = \text{BN}(\text{AAP}(f_i^R)), \tag{8}$$

$$O = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V, \tag{9}$$

$$\tilde{f}_{e,i}^F = \text{Concat}(O, f_{e,i}^F), \tag{10}$$

where $f_{e,i}^F$ denotes the output feature of the i-th stage of the fusion network encoder and f_i^R denotes the reconstructed feature of the i-th IRM. BN$(\cdot)$ represents the batch normalization operation, which stabilizes training and accelerates convergence. AAP$(\cdot)$ denotes the adaptive average pooling operation, which dynamically adjusts the size of the feature map. Concat$(\cdot)$ represents feature concatenation operation, used to preserve more information. Additionally, softmax is used to normalize scores, ensuring that the sum of weights is 1, which is typically applied in attention mechanisms.

3.3 Feature Fusion Module

As shown in Figure 5, the feature fusion module (FFM) is mainly composed of convolutional layers and the activation function is ReLU, which is designed to enhance the fusion network decoder's ability to understand semantic information. Specifically, the reconstructed feature f_i^R is firstly encoded with double 3×3 convolutions and concatenated with the fusion network decoder feature $f_{d,n-i+1}^F$. The concatenated features are further deepened and fused through two 3×3 convolutional layers. This process can be represented as:

$$f_{\text{concat}} = \text{Concat}(\text{Conv}_{3\times3}(\text{Conv}_{3\times3}(f_i^R)), f_{d,n-i+1}^F), \tag{11}$$

$$f_{\text{deepened}} = \text{Conv}_{3\times3}(\text{Conv}_{3\times3}(f_{\text{concat}})), \tag{12}$$

where $\text{Conv}_{3\times3}$ denotes the operation of a 3×3 convolutional layer, mainly used for feature extraction of input features. Subsequently, the deepened features are concatenated with the original $f_{d,n-i+1}^F$ again, and finally the number of channels is adjusted through a 1×1 convolutional layer and output:

$$\tilde{f}_{d,n-i+1}^F = \text{Conv}_{1\times1}(\text{Concat}(f_{\text{deepened}}, f_{d,n-i+1}^F)), \tag{13}$$

where $\text{Conv}_{1\times1}$ denotes a 1×1 convolutional layer operation, mainly used to adjust the number of feature channels.

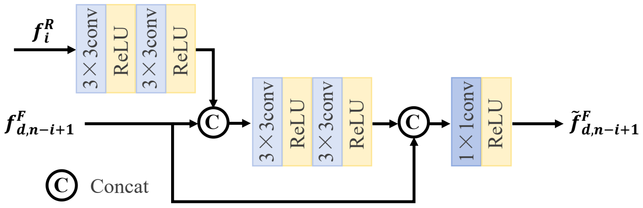


Figure 5. The structure of the feature fusion module.

3.4 Loss function

To optimize the proposed model, we design a loss function. It converts semantic-level segmentation features into pixel-level features via IRM, bridging the feature gap between segmentation and fusion, enabling their interaction, and improving fusion quality. Specifically, we jointly train fusion, segmentation and reconstruction tasks. Therefore, the designed loss function can be expressed as:

$$L_{\text{total}} = L_f + L_s + L_{\text{rec}}, \tag{14}$$

where L_f and L_s represent the image fusion loss and segmentation loss, respectively. And L_{rec} represents the reconstruction loss of segmentation semantic features.

In the fusion stage, the fusion loss is defined as:

$$L_f = (1 - \text{SSIM}(I_F, I_{\text{ir}})) + (1 - \text{SSIM}(I_F, I_{\text{vis}})), \tag{15}$$

where SSIM [26] represents the structural similarity index, which is used to evaluate the difference between the fusion result I_F and the source images I_{ir} and I_{vis}^Y.

The reconstruction loss mainly evaluates the similarity between the reconstructed image and the original infrared and visible images. The L_{rec} is defined as:

$$L_{\text{rec}} = \sum_{i=1}^{n} L_{\text{rec},i}, \tag{16}$$

$$L_{\text{res}i} = (1 - \text{SSIM}(I_{\text{res}i}, I_{\text{ir}})) + (1 - \text{SSIM}(I_{\text{res}i}, I_{\text{vis}})), \tag{17}$$

where $L_{\text{rec},i}$ represents the reconstruction loss for the i-th image reconstruction module, and $I_{\text{res}i}$ is the feature output by the i-th image reconstruction module.

In the segmentation stage, the segmentation loss function L_s is composed of the cross-entropy (CE) loss [29] and the dice coefficient loss [27, 28]:

$$L_{\text{s}} = L_{\text{ce}} + L_{\text{dice}}, \tag{18}$$

where CE loss is used to measure the difference between the predicted probabilities and the true labels, providing a measure of classification accuracy. The dice coefficient loss, on the other hand, evaluates the similarity between the predicted and true segmentations, making it particularly useful for tasks where the goal is to match the predicted segmentation with the true segmentation.

4 Experiments

4.1 Experimental configurations

The model is implemented with PyTorch on GTX 2080TI GPU. During training, we employ the Adam optimizer with a learning rate of 0.0001. The two momentum values of the Adam optimizer are set to 0.9 and 0.999, respectively. The batch size is set to 2, and the number of epochs is set to 50. We train the model on the WHU [44] and Potsdam [45]. The Potsdam dataset provides detailed information about urban environments, mainly including 6 categories: Impervious surfaces, Buildings, Low vegetation, Trees, Cars, and Clutter. It is divided into a training set of 10,830 images and a test set of 2,527 images. The WHU dataset describes the scenario of land, covering 7 categories: Farmland, City, Village, Water, Forest, Road, and Others [46]. It is divided into a training set of 17,280 images and a test set of 4,320 images. Before training, we preprocess the data by cropping the images into patches of size 320×320.

In addition, for quantitative evaluation, four metrics are selected to objectively evaluate the fusion performance, including spatial frequency (SF) [40], average gradient (AG) [41], the sum of the correlations of differences (SCD) [10], and visual information fidelity (VIF) [12]. SF measures the richness of detail information in the image. AG reflects the clarity of the image. SCD reflects the degree of correlation between the information transferred to the fused image and the corresponding source image. VIF measures the degree of visual information preservation of the fused image relative to the source image from the perspective of human visual perception. The larger the SF, AG, SCD and VIF of the fusion algorithm, the better the fusion performance.

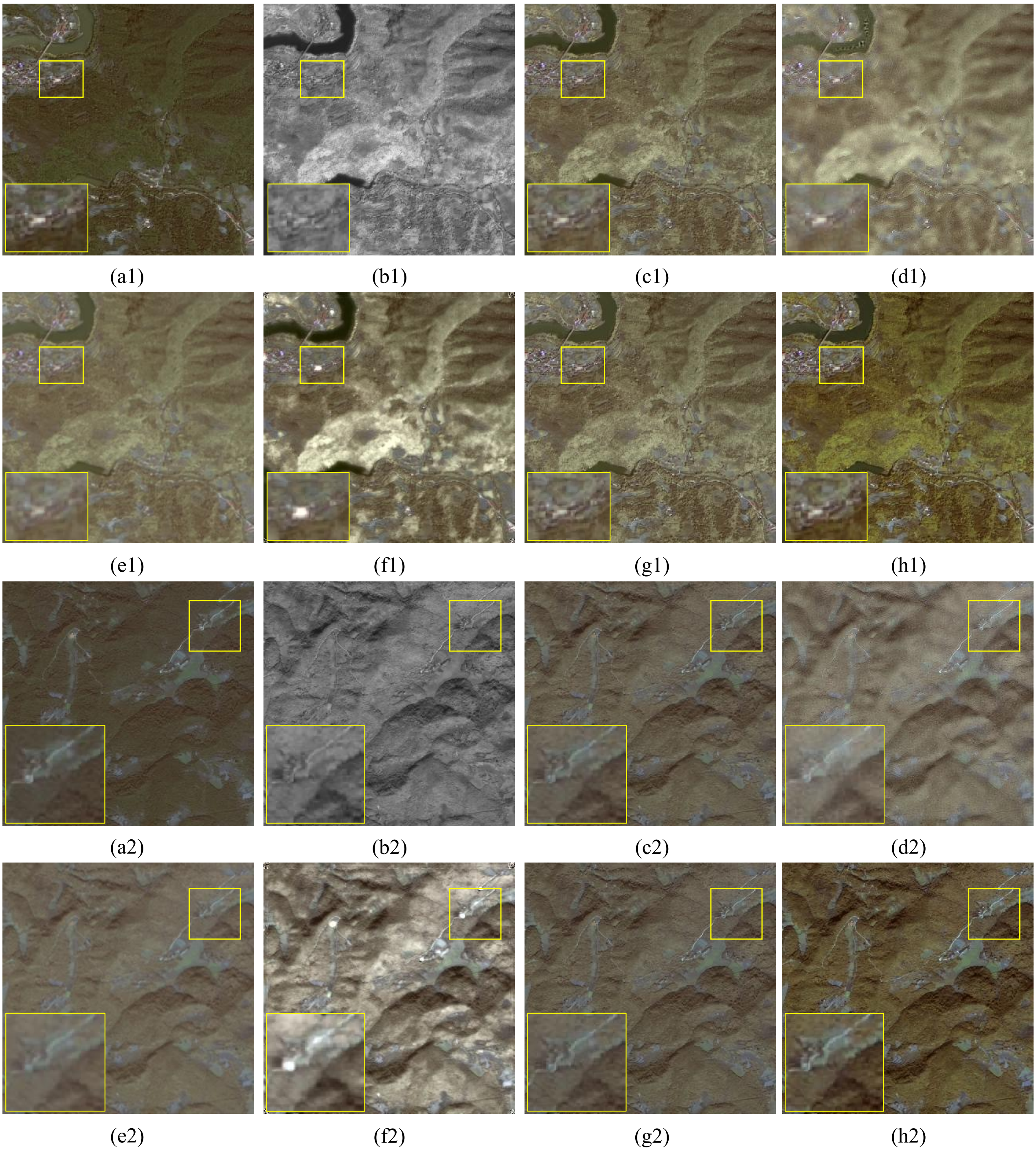
4.2 Results and analysis

In this section, we conduct subjective qualitative and objective quantitative experiments on the WHU and Potsdam datasets to evaluate the performance and advantages of our proposed fusion method. We select five state-of-the-art methods, including UMFusion [25], YDTR [19], Tardal [17], ITFuse [42] and LiMFusion [43], to compare with the proposed model. Next, we conduct a detailed analysis of the fusion results obtained by these methods on the WHU and Potsdam datasets from both subjective and objective dimensions.

4.2.1 Experimental results on the WHU dataset

First, we qualitatively compare the proposed method with five comparison methods. We select two representative infrared and visible images from the WHU dataset for subjective evaluation, as shown in Figure 6. In the picture, in order to visually compare the fusion effects of different methods, we mark the comparison area with a yellow box, and enlarge the details of the corresponding area and display it in the lower left corner of the image. From the visualization results, it can be observed that the fusion images generated by ITFuse and LiMFusion exhibit lower clarity, with blurred edges and loss of some detail information. Although UMFusion and YDTR have fused infrared and visible information to a certain extent, there is still room for improvement in detail preservation and clarity. Tardal has a high contrast and highlights the infrared thermal radiation target well, but there are local artifacts, and the preservation of texture and edge detail information is not as good as the proposed method. In contrast, the proposed method performs better in integrating the complementary features of infrared and visible images, and the generated fused images have higher clarity, can well preserve edge and texture detail information, and are more in line with the characteristics of the human visual system. Therefore, in the qualitative comparison of infrared and visible image fusion methods, the visualization results of the proposed method outperform those of existing state-of-the-art methods.

In addition, to comprehensively evaluate the performance of the proposed method and five comparison methods, four metrics, SF, AG, SCD and VIF are used to quantitatively analyze the fused image.


(a1) (b1) (c1) (d1)

(e1) (f1) (g1) (h1)

(a2) (b2) (c2) (d2)

(e2) (f2) (g2) (h2)

Figure 6. Visual comparison of our method with five SOTA fusion methods on the WHU dataset. (a1-h1) and (a2-h2) represent visible image, infrared image, UMFusion, LiMFusion, ITFuse, Tardal, YDTR and our proposed model, respectively.

As shown in Table 1, the average comparison results of the four metrics of the proposed method and other comparison methods on the WHU test set are shown, where the optimal value of each metric is marked in red and the suboptimal value is marked in blue. Obviously, the proposed method is higher than the existing comparison methods in the three evaluation metrics of AG, SCD and VIF, which shows that the fused image generated by the proposed method has the best performance in clarity and retains more feature information of the source image, with better visual performance. Although the proposed method is not the best in the metric of spatial frequency (SF), it is second only to LiMFusion, and the gap is not large, which shows that the fused image generated by the proposed method contains richer texture and edge

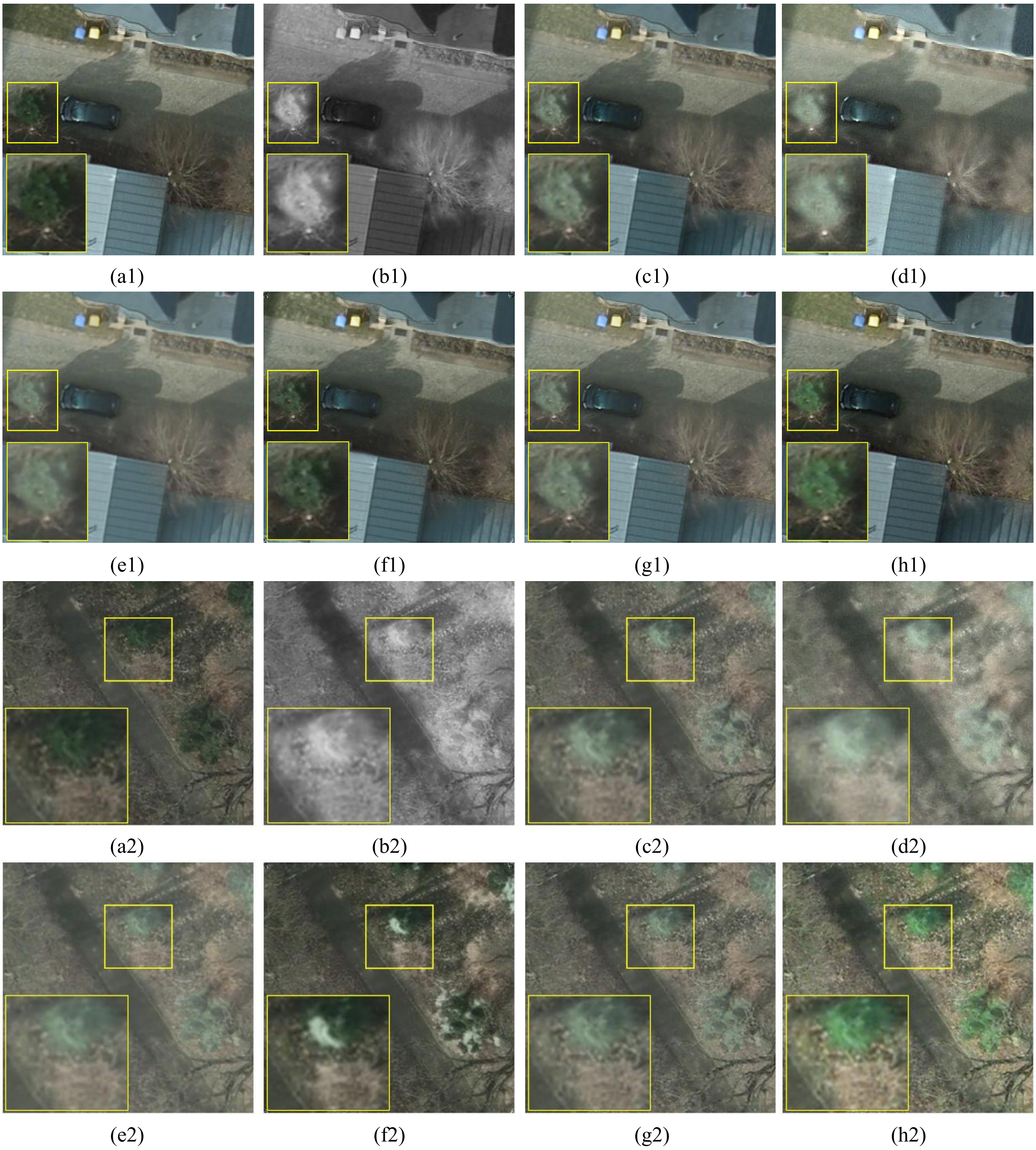


Figure 7. Visual comparison of our method with five SOTA fusion methods on the Potsdam dataset. (a1-h1) and (a2-h2) represent visible image, infrared image, UMFusion, LiMFusion, ITFuse, Tardal, YDTR, and our proposed model, respectively.

detail information.

4.2.2 *Experimental results on the Potsdam dataset*

We further conduct experiments on the Potsdam dataset and conduct qualitative and quantitative analysis of the experimental results to demonstrate the effectiveness and superiority of the proposed method on different datasets. Figure 7 shows subjective visualization results of two sets of infrared and visible images. From the experimental results, it can be seen that the fused images generated by LiMFusion and ITFuse have low clarity, some detail loss, and low overall visual quality. The fused images generated by UMFusion and YDTR are not clear enough, the contrast is relatively low, and some texture detail information is lost. The fused images generated by Tardal can retain

Table 1. Average quality metrics of different methods on the WHU dataset. The optimal result is highlighted in red, and the sub-optimal result is highlighted in blue.

Methods	SF	AG	SCD	VIF
YDTR [19]	14.365	5.01	0.829	1.015
Tardal [17]	15.465	5.867	1.033	0.796
UMFusion [25]	13.4735	4.8998	0.8378	1.0329
ITFuse [42]	6.6021	2.8736	0.2886	0.6699
LiMFusion [43]	17.4743	5.7829	0.6069	0.5147
Ours	16.8724	5.9824	1.3197	1.0969

texture detail information, but the visual effects are poor. In contrast, the proposed method offers better visual effects, retains more source image information, has higher contrast, and better meets human visual system needs.

In addition, Table 2 shows the objective comparison results of different fusion methods on the Potsdam test set. The proposed method has achieved optimal or near-optimal values in most metrics. Specifically, the proposed method performs best in the three metrics of AG, SCD and VIF, which shows that the fused image generated by the proposed method not only has the highest clarity but also can more effectively fuse the key feature information in the source image into the final result, with good fusion quality, which is more in line with the human visual system. In terms of SF, the performance of the proposed method is second only to LiMFusion, which shows that the fused image generated by the proposed method contains relatively rich edge and texture detail information.

Table 2. Average quality metrics of different methods on the Potsdam dataset. The optimal result is highlighted in red, and the sub-optimal result is highlighted in blue.

Methods	SF	AG	SCD	VIF
YDTR [19]	10.180	3.608	0.310	1.396
Tardal [17]	9.915	3.409	0.524	1.238
UMFusion [25]	8.7409	3.3025	0.5245	1.2752
ITFuse [42]	6.1962	2.4580	0.1746	1.0358
LiMFusion [43]	11.4870	4.0733	0.8851	0.7541
Ours	11.3001	4.1599	1.1660	1.5861

In summary, the experimental results on both WHU and Potsdam datasets show that the proposed method exhibits superior performance in infrared and visible image fusion compared with five state-of-the-art methods.

4.3 Ablation study

4.3.1 Effect of stage-interactive network

As shown in Figure 3, we introduce stage-interactive networks between the encoder stages of the segmentation network and the fusion network. The stage-interactive network mainly includes an image reconstruction module, a cross attention module, and a feature fusion module. The stage-interactive network is responsible for bridging the feature gap between segmentation and fusion tasks, thereby achieving feature interaction between segmentation and fusion tasks and improving the performance of fusion tasks. In this study, a total of three stages of feature interaction are used. This section aims to explore the impact of the number of stage-interactive networks on model performance. Table 3 lists in detail the quantitative evaluation results of different numbers of stage-interactive networks on the WHU dataset.

Table 3. Average quality metrics of different numbers of stage-interactive networks. The optimal result is bolded.

Number	SF	AG	SCD	VIF
1	16.8547	5.9754	1.3188	1.0966
2	16.8693	5.9798	1.3182	1.0971
3	**16.8724**	**5.9824**	**1.3197**	**1.0969**

As can be seen from Table 3, as the number of stage-interactive networks increases, the performance of the fusion task is generally improved, which indicates that a greater number of stage-interactive networks can more effectively enhance the interaction between the features of the fusion and segmentation tasks, so that the fusion network can better utilize the semantic information of the segmentation network. When the number of stage-interactive network is 3, most metrics reach the optimal or near-optimal values, which shows that the fusion effect is better at this time, the generated fusion image has higher clarity, richer edge and texture detail information, and has more advantages in meeting the needs of the human visual system.

4.3.2 Effect of image reconstruction module

The role of the image reconstruction module (IRM) is to align the target-level features of the segmentation network with the pixel-level features of the image fusion task, thereby bridging the feature gap between image fusion and image segmentation. In this section, to fully verify the effectiveness of the image reconstruction module, we design a series

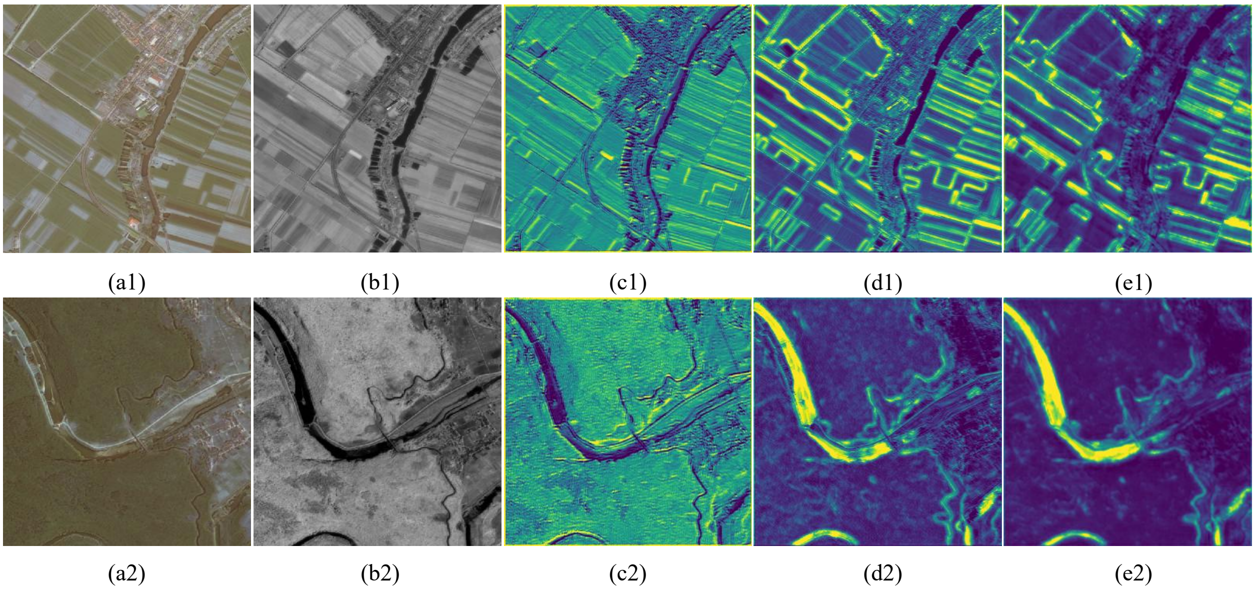


Figure 8. Visual comparison of feature alignment results before and after IRM processing. (a1-e1) and (a2-e2) represent visible image, infrared image, segmentation feature, reconstructed feature, and fusion feature.

of comparative experiments. Firstly, for the fusion model containing two stage-interactive networks, we conduct two experiments: one retains the image reconstruction module and the other removes the module while keeping other structures unchanged, in order to observe the specific impact of the image reconstruction module on the quality of the fused image.

In addition, to further explore the performance of the image reconstruction module under different configurations, we also add a set of experiments to compare the performance difference between retaining and deleting the image reconstruction module in a fusion model that includes a three stage-interactive networks. The experimental results of the objective evaluation metrics are shown in Table 4.

Table 4. Effect study of image reconstruction module (IRM). The optimal result is bolded. And w/ means with, w/o means without.

Setting	SF	AG	SCD	VIF
Two-SINets (w/ IRM)	**16.8693**	**5.9798**	**1.3182**	**1.0971**
Two-SINets (w/o IRM)	16.8419	5.9713	1.3179	1.0967
Three-SINets (w/ IRM)	**16.8724**	**5.9824**	**1.3197**	**1.0969**
Three-SINets (w/o IRM)	16.8317	5.9695	1.3189	1.0964

As shown in Table 4, by comparing the results of the first two experiments, it can be seen that the setting including the image reconstruction module is better than the setting without the module in all metrics, proving that the image reconstruction module can improve the quality of the fused image. Similarly, the comparison of the results of the last two experiments also verifies this point, further proving the effect of the image reconstruction module.

Moreover, we validate the alignment effect of the image reconstruction module by visualizing intermediate features (comparing segmentation features, reconstructed features, and fusion features). As shown in Figure 8, the segmentation features before alignment lack pixel-level detail information (e.g., the edge contours of farmland are blurred). In contrast, after adding the image reconstruction module, the boundaries between farmland and water are clearly presented, and semantic-level features are reconstructed into pixel-level features suitable for image fusion, achieving accurate feature alignment.

4.3.3 *Effect of cross attention module*

The cross attention module (CAM) is responsible for promoting the interaction of features between image segmentation and image fusion, thereby helping the fusion task to better fuse salient targets during the fusion process to improve the quality of the fused image. In this section, in order to fully verify the effect of the cross attention module, we design two sets of comparative experiments. Firstly, for the fusion model containing two stage-interactive networks, we conduct two experiments: one retains the cross attention module, and the other replaces the cross attention module with a simple feature addition operation. Secondly, we also compare the performance differences when retaining the cross attention module and replacing the cross attention module with a feature addition operation in the context of the three stage-interactive networks. The experimental results are shown in Table 5.

Table 5. Effect study of cross attention module (CAM). The optimal result is bolded. And w/ means with, w/o means without.

Setting	SF	AG	SCD	VIF
Two-SINets (w/ CAM)	16.8693	**5.9798**	**1.3182**	**1.0971**
Two-SINets (w/o CAM)	**16.8699**	5.9786	1.317	1.0966
Three-SINets (w/ CAM)	**16.8724**	**5.9824**	1.3197	**1.0969**
Three-SINets (w/o CAM)	16.8684	5.9815	**1.3206**	1.0969

As shown in Table 5, the results show that whether it is a two stage-interactive or a three stage-interactive, the introduction of the cross attention module is more helpful in improving the quality of the fused image than the simple feature addition operation.

To further validate the effectiveness of the CAM module, we visualize the distribution of attention weights using heatmaps (see Figure 9). As observed from the heatmaps, attention weights are predominantly focused on target regions (e.g., forest), which enables the segmentation task to effectively assist the fusion task and thereby enhancing the quality of image fusion.

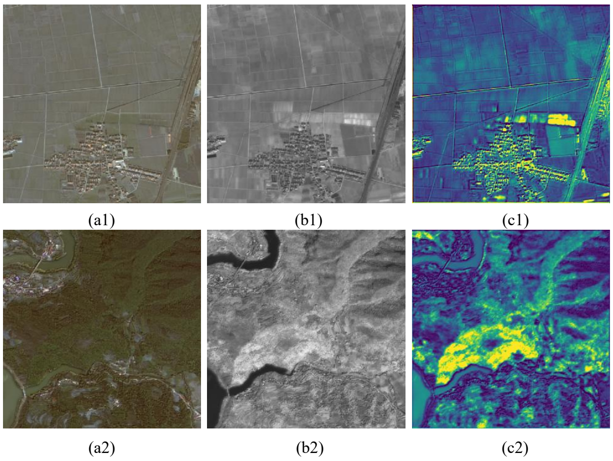


Figure 9. Qualitative analysis of attention weights. (a1-c1) and (a2-c2) represent visible image, infrared image, and heatmap.

5 Conclusion

To address the difference in feature representation between image fusion and image segmentation, this paper proposes self-supervised feature alignment for infrared and visible image fusion. The innovation of this method lies in the design of an image reconstruction module, which aligns the target-level features extracted by the segmentation network with the pixel-level features extracted by the fusion network through a self-supervised method, effectively bridging the feature gap between image fusion and image segmentation. In addition, the cross attention module is introduced to promote feature interaction between the two tasks, thereby achieving efficient collaboration between image segmentation and image fusion tasks. Finally, the performance advantages of the proposed method are demonstrated through comprehensive qualitative and quantitative analysis on the WHU and Potsdam datasets. However, the performance of the proposed method depends on the supervision information provided by the segmentation task. In future work, we will explore how to mine semantic information to improve fusion quality without segmentation supervision.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported by the National Natural Science Foundation of China under Grant 62522105.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Das, S., & Zhang, Y. (2000). Color night vision for navigation and surveillance. *Transportation Research Record*, *1708*(1), 40–46. [CrossRef]

[2] Paramanandham, N., & Rajendiran, K. (2018). Multi sensor image fusion for surveillance applications using hybrid image fusion algorithm. *Multimedia Tools and Applications*, *77*(10), 12405-12436. [CrossRef]

[3] Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U., & Yu, Y. (2023). Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, *90*, 185-217. [CrossRef]

[4] Qi, J., Liang, T., Liu, W., Li, Y., & Jin, Y. (2024). A Generative-Based Image Fusion Strategy for Visible-Infrared Person Re-Identification. *IEEE*

Transactions on Circuits and Systems for Video Technology, 34(1), 518–533. [CrossRef]

[5] Li, H., Ding, W., Cao, X., & Liu, C. (2017). Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing. *Remote Sensing, 9*(5), 441. [CrossRef]

[6] Ruan, Z., Wan, J., Xiao, G., Tang, Z., & Ma, J. (2024). Semantic attention-based heterogeneous feature aggregation network for image fusion. Pattern Recognition, 155, 110728. [CrossRef]

[7] Xu, X., Wang, S., Wang, Z., Zhang, X., & Hu, R. (2021). Exploring image enhancement for salient object detection in low light images. *ACM transactions on multimedia computing, communications, and applications (TOMM), 17*(1s), 1-19. [CrossRef]

[8] Gao, Y., Ma, S., & Liu, J. (2023). DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(2), 549-561. [CrossRef]

[9] Liu, R., Ma, L., Ma, T., Fan, X., & Luo, Z. (2023). Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 5953-5969. [CrossRef]

[10] Aslantas, V., & Bendes, E. (2015). A new image quality metric for image fusion: The sum of the correlations of differences. *AEU - International Journal of Electronics and Communications, 69*(12), 1890-1896. [CrossRef]

[11] Jian, L., Yang, X., Liu, Z., Jeon, G., Gao, M., & Chisholm, D. (2021). SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-15. [CrossRef]

[12] Han, Y., Cai, Y., Cao, Y., & Xu, X. (2013). A new image fusion performance metric based on visual information fidelity. *Information Fusion, 14*(2), 127-135. [CrossRef]

[13] Li, H., & Wu, X. J. (2018). DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing, 28*(5), 2614-2623. [CrossRef]

[14] Li, H., Wu, X.-J., & Durrani, T. (2020). NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement, 69*(12), 9645-9656. [CrossRef]

[15] Ma, J., Yu, W., Liang, P., Li, C., & Jiang, J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion, 48*, 11-26. [CrossRef]

[16] Ma, J., Xu, H., Jiang, J., Mei, X., & Zhang, X.-P. (2020). DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing, 29*, 4980-4995. [CrossRef]

[17] Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., & Luo, Z. (2022). 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* (pp. 5792-5801). [CrossRef]

[18] Tang, L., Yuan, J., & Ma, J. (2022). 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. [CrossRef]

[19] Tang, W., He, F., & Liu, Y. (2023). YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Transactions on Multimedia, 25*, 5413-5428. [CrossRef]

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

[21] Li, J., Huo, H., Li, C., Wang, R., Sui, C., & Liu, Z. (2021). Multigrained attention network for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-12. [CrossRef]

[22] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., & Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion, 54*, 99-118. [CrossRef]

[23] Ma, J., Tang, L., Xu, M., Zhang, H., & Xiao, G. (2021). STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-13. [CrossRef]

[24] Tang, L., Yuan, J., Zhang, H., Jiang, X., & Ma, J. (2022). PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion, 83*, 79-92. [CrossRef]

[25] Wang, D., Liu, J., Fan, X., & Liu, R. (2022). Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*.

[26] Wang, Z., Bovik, A.C., Sheikh, H.R., & Simoncelli, E.P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600-612. [CrossRef]

[27] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565-571). [CrossRef]

[28] Crum, W.R., Camara, O., & Hill, D.L.G. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging, 25*(11), 1451-1461. [CrossRef]

[29] Kline, D.M., & Berardi, V.L. (2005). Revisiting squared-error and cross-entropy functions for training

neural network classifiers. *Neural Computing and Applications, 14*, 310–318. [CrossRef]

[30] Zhang, X. (2021). Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4819-4838. [CrossRef]

[31] Shelhamer, E., Long, J., & Darrell, T. (2016). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 640-651. [CrossRef]

[32] Xu, H., Wang, X., & Ma, J. (2021). DRF: Disentangled Representation for Visible and Infrared Image Fusion. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-13. [CrossRef]

[33] Li, J., Huo, H., Li, C., Wang, R., & Feng, Q. (2021). AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. *IEEE Transactions on Multimedia, 23*, 1383-1396. [CrossRef]

[34] Huang, S., Song, Z., Yang, Y., Wan, W., & Kong, X. (2023). MAGAN: Multiattention Generative Adversarial Network for Infrared and Visible Image Fusion. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-14. [CrossRef]

[35] Fu, Y., Liu, Z., Peng, J., Gupta, R., & Zhang, D. (2025). GANSD: A generative adversarial network based on saliency detection for infrared and visible image fusion. *Image and Vision Computing, 154*, 105410. [CrossRef]

[36] Hu, X., Liu, Y., & Yang, F. (2024). PFCFuse: A Poolformer and CNN Fusion Network for Infrared-Visible Image Fusion. *IEEE Transactions on Instrumentation and Measurement, 73*, 1-14. [CrossRef]

[37] Lu, Q., Zhang, H., & Yin, L. (2025). Infrared and visible image fusion via dual encoder based on dense connection. *Pattern Recognition, 163*, 111476. [CrossRef]

[38] Wang, W., Deng, L.-J., Ran, R., & Vivone, G. (2024). A General Paradigm with Detail-Preserving Conditional Invertible Network for Image Fusion. *International Journal of Computer Vision, 132*(4), 1029–1054. [CrossRef]

[39] Liu, R., Jiang, Z., Yang, S., & Fan, X. (2022). Twin Adversarial Contrastive Learning for Underwater Image Enhancement and Beyond. *IEEE Transactions on Image Processing, 31*, 4922–4936. [CrossRef]

[40] Zheng, Y., Essock, E. A., Hansen, B. C., & Haun, A. M. (2007). A new metric based on extended spatial frequency and its application to DWT based fusion algorithms. *Information Fusion, 8*(2), 177-192. [CrossRef]

[41] Cui, G., Feng, H., Xu, Z., Li, Q., & Chen, Y. (2015). Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications, 341*, 199-209. [CrossRef]

[42] Tang, W., He, F., & Liu, Y. (2024). ITFuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition, 156*, 110822. [CrossRef]

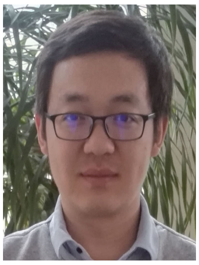
[43] Qian, Y., Tang, H., Liu, G., Xing, M., Xiao, G., & Bavirisetti, D. P. (2024). LiMFusion: Infrared and visible image fusion via local information measurement. Optics and Lasers in Engineering, 181, 108435. [CrossRef]

[44] Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., & Zhang, L. (2022). MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation, 106*, 102638. [CrossRef]

[45] Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Bnitez, S., & Breitkopf, U. (2020). International society for photogrammetry and remote sensing, 2d semantic labeling contest. Accessed: Oct,29.

[46] Zhao, W., Cui, H., Wang, H., He, Y., & Lu, H. (2025). FreeFusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(9), 8040-8056. [CrossRef]

Weitao Qiu received the B.S. degree from Hebei University of Technology (HEBUT), Tianjin, China, in 2022. She is currently pursuing the M.S. degree at Dalian University of Technology (DUT), Dalian, China. Her research interests include image fusion. (Email: qiuweit7@163.com)



Wenda Zhao received the B.S. degree from Jilin University, Changchun, China, in 2011, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2016. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology (DUT), Dalian, China. His research interests include deep learning and computer vision tasks. (Email: zhaowenda@dlut.edu.cn)

Haipeng Wang received the Ph.D. degree from Naval Aviation University, Yantai, China, in 2012. He is currently a Professor with Unit 92728 of PLA. His research interests include the general area of intelligent perception and fusion, and big data technology and applications.