Contents lists available at ScienceDirect

Expert Systems With Applications

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

SFINet: A semantic feature interactive learning network for full-time infrared and visible image fusion

Wenhao Song [a,1], Qilei Li [a,b,1], Mingliang Gao [a,*], Abdellah Chehri [c], Gwanggil Jeon [d]

[a] *School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, Shandong, China*
[b] *School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom*
[c] *Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, K7K 7B4, Canada*
[d] *Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, South Korea*

ARTICLE INFO

Keywords:
Image fusion
Deep learning
Semantic information
Attention mechanism

ABSTRACT

Infrared and visible image fusion aims to combine data from various source images to generate a high-quality image. Nevertheless, numerous fusion methods often prioritize visual quality above semantic information. To address this problem, we present a Semantic Feature Interactive Learning Network (SFINet) for full-time infrared and visible images. The SFINet encompasses an image fusion network and an image segmentation network through a Semantic Feature Interaction (SFI) module. The image fusion network employs Multi-scale Feature Extraction (MFE) modules to capture global and local information at multiple scales. Meanwhile, it performs an adaptive fusion of complementary information using a Dual Attention Feature Fusion (DAFF) module. The image segmentation network guides the image fusion network using the SFI module for semantic feature interaction. Comparative results prove that the proposed method is superior to state-of-the-art (SOTA) models in image fusion and semantic segmentation tasks. The code is available at https://github.com/songwenhao123/SFINet.

1. Introduction

A single image taken by the same device falls short of fully reflecting the scene due to technical and environmental factors (Karim et al., 2023). Although infrared (IR) sensors excel at revealing salient targets by detecting thermal radiation, they suffer from low texture and high noise. On the other hand, visible (VIS) sensors can capture rich textures and structures via reflected light, yet they are vulnerable to environmental factors like illumination and occlusion. Infrared and visible image fusion (IVIF) is a technique that capitalizes on the strengths of source images. It can yield a fused image that highlights important targets and displays detailed information.

Recently, IVIF has attracted significant interest from the academic community, and many image fusion methods have been proposed. The IVIF methods are broadly divided into traditional methods (Zou & Yang, 2023) and deep learning (DL)-based methods (Song, Zhai et al., 2024). Traditional methods decompose an image into a feature space and then fuse the features based on manually crafted fusion rules. Finally, the fused image is generated by reconstructing the fused features (Ma, Ma et al., 2019). Nevertheless, handcrafted fusion rules often fall short of meeting the requirements of downstream tasks and exhibit significant limitations (Zhang, Xu et al., 2021). The evolution of deep learning has consequently spurred considerable interest in image fusion. Three categories of DL-based image fusion methods are auto-encoder (AE)-based method (Jian et al., 2020), convolutional neural network (CNN)-based method (Song, Gao et al., 2024), and generative adversarial network (GAN)-based method (Gao et al., 2023). AE-based methods utilize auto-encoders for feature extraction and image reconstruction, which apply manually designed fusion rules to integrate the features. CNN-based methods combine feature extraction, fusion, and image reconstruction through intricately designed networks and loss functions. GAN-based methods achieve an unsupervised image fusion by building an adversarial between the generator and the discriminator.

IVIF has its way of assisting semantic segmentation tasks. The benefit of image fusion for the segmentation network is illustrated in Fig. 1. It shows that the IR image enables the network to identify prominent targets like persons and cars but overlooks crucial information like roadblocks. Although the segmentation result of cars and roadblocks based on the VIS image is fair, the effect of person segmentation is

* Corresponding author.
E-mail addresses: 18110403087@stumail.sdut.edu.cn (W. Song), q.li@qmul.ac.uk (Q. Li), mlgao@sdut.edu.cn (M. Gao), abdellah.chehri@rmc-cmr.ca (A. Chehri), gjeon@inu.ac.kr (G. Jeon).

[1] The two authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2024.125472
Received 10 January 2024; Received in revised form 25 September 2024; Accepted 27 September 2024
Available online 5 October 2024

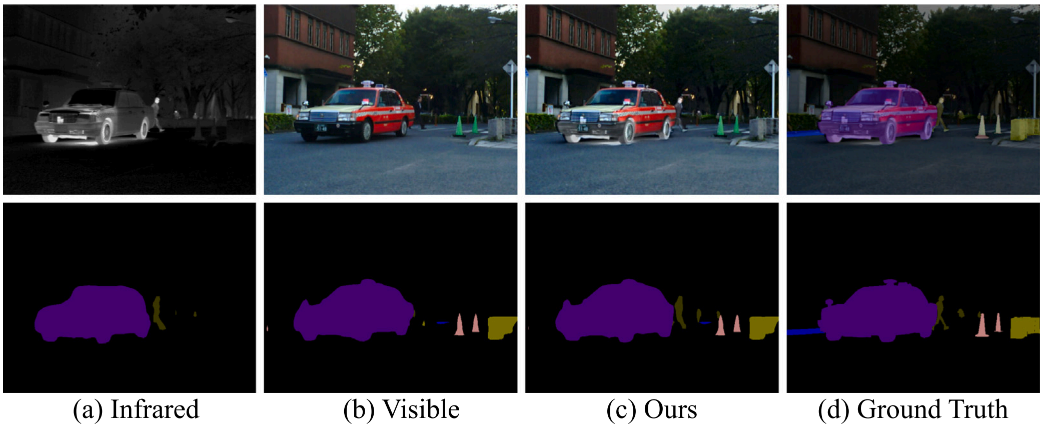


Fig. 1. Sematic segmentation results based on IR image, VIS image, and fused image.

unsatisfactory. In comparison, the segmentation result based on the fused image is close to the ground truth, with the persons, cars, and roadblocks being well segmented.

The existing semantic segmentation task-driven IVIF models can be divided into semantic loss-driven networks, *e.g.*, SeAFusion (Tang, Yuan, Ma, 2022) and SuperFusion (Tang, Deng et al., 2022), and semantic feature-driven networks, *e.g.*, PSFusion (Tang, Zhang et al., 2023). The architectures of the state-of-the-art semantic segmentation-driven image fusion models and the proposed model are depicted in Fig. 2. For the semantic loss-driven networks, the fused image generated by the fusion network is input into the segmentation network to obtain semantic representation, and the semantic loss guidance directs the fusion network to focus on more semantic information.

Since semantic segmentation networks cannot achieve the interaction of semantic features, the fusion networks only focus on pixel-level features and cannot extract deep semantic-level information. Therefore, the semantic feature-driven networks were proposed. For the semantic feature-driven networks, the shallow features and deep features are first extracted through the feature extraction backbone network. The shallow features are input into the fusion network, while the deep features are fed into the segmentation network. Then, some features from the segmentation network are fed back to the fusion network to compensate for the semantic features. Although the fusion network can be guided by the semantic segmentation network, the parameters and the complexity of the model are increased.

In this paper, we designed a Semantic Feature Interactive Learning Network (SFINet) considering the advantages of semantic loss-driven models and semantic feature-driven models. Specifically, a Semantic Feature Interaction (SFI) module is built to compensate for the semantic feature, and a Multi-scale Feature Extraction (MFE) module is introduced to extract semantic information and multi-scale information from the source image. Meanwhile, a Dual Attention Feature Fusion (DAFF) module is developed to combine to learn the correlations between different modal features jointly. Overall, the contributions of this work can be summarized as follows.

- We formulate a comprehensive framework that simultaneously handles cross-modality image fusion and semantic segmentation. The proposed framework demonstrates outstanding performance due to the tailored semantic feature interaction module.
- We derive a multi-scale feature extraction module that captures information at hierarchical levels, which effectively balances high-frequency local details and low-frequency global context and prompts the learning of discriminative representation.
- We design a dual-attention feature fusion module to explore the correlation among multimodal features to ensure unbiased information fusion among different modalities, which enhances the versatility and superiority of the features for both fusion and segmentation tasks.
- We propose a semantic feature interaction module to augment the semantic information of fused features. This is achieved through facilitating interactions between features from the segmentation network and those from the fusion network.

The rest of the paper is structured as follows. Section 2 introduces recent works related to the proposed method. Section 3 describes the proposed SFINet in detail. Section 4 presents the experimental results and the ablation studies. Section 5 concludes this work.

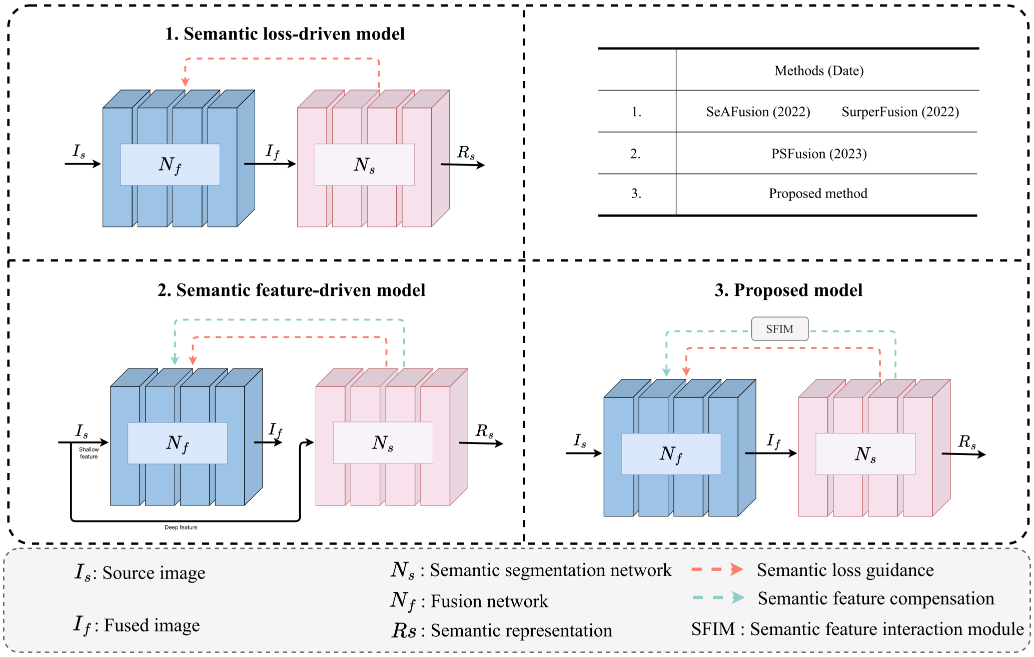
2. Related work

2.1. Traditional image fusion methods

Traditional IVIF methods rely on two key steps, namely, feature extraction and fusion. These methods can be divided into four categories, *i.e.*, multi-scale transform methods (Zhou et al., 2016), sparse representation methods (Zhang et al., 2018), subspace-based methods (Mitchell, 2010), and hybrid methods (Gan et al., 2015). Multi-scale transform methods break down the source images into different scales and then combine them by taking into account specific measurements of activity levels. For example, Yan et al. (2015) proposed an IVIF method that utilizes spectral graph wavelet transform and a bilateral filter. Sparse representation methods employ an overcompleted dictionary to represent images as sparse coefficients and then fuse them based on specific sparsity criteria. For instance, Wu et al. (2020) developed a method based on convolutional sparse representation, which can preserve spatial consistency and distinct features from infrared and visible images. Subspace-based methods reduce the dimensionality of images and capture their intrinsic structures. For instance, Fu et al. (2016) introduced a technique based on joint convolutional sparse representation, which can preserve spatial consistency and distinct features from the source images. Hybrid methods integrate the strengths of the above approaches to attain enhanced fusion performance. For instance, Gan et al. (2015) employed multi-scale decomposition and guided filters for IVIF. This method improves visual quality and reduces artifacts in the fused image by incorporating saliency maps and weighting maps.

2.2. DL-based image fusion methods

Deep learning has demonstrated exceptional performance across diverse vision tasks and has been widely applied in image fusion tasks. The primary strength of deep learning lies in its capacity to autonomously learn features from data and eliminate the need for handcrafted rules or transformations. The DL-based image fusion methods can be categorized into Auto-encoder (AE)-based fusion image methods, CNN-based fusion image methods, and GAN-based fusion image methods.



	Methods (Date)	
1.	SeAFusion (2022)	SurperFusion (2022)
2.	PSFusion (2023)	
3.	Proposed method	

Fig. 2. Architectures of the existing advanced semantic segmentation-driven image fusion models.

2.2.1. *AE-based image fusion methods*

AE-based methods leverage an auto-encoder to extract the feature and reconstruct the fused image. This architecture comprises an encoder and a decoder. The encoder maps the source image to a latent feature space while the decoder reconstructs the output image from the latent features. For example, Li and Wu (2018) proposed an AE-based IVIF method. This approach incorporates a fusion layer and the dense block to extract and integrate features from the source images. Additionally, a decoder reconstructs the fused image. Xu, Zhang et al. (2021) employed an encoder to evaluate the importance of feature maps and fuse them according to classification saliency maps. This unsupervised method eliminates the need for hand-crafted fusion rules. Xu, Wang et al. (2021) utilized disentangled representation to decompose the source images into scene-related and sensor-related components. The fusion is performed on these components using different strategies and a pre-trained generator.

2.2.2. *CNN-based image fusion methods*

CNN-based fusion methods implicitly perform feature extraction, fusion, and image reconstruction with the complex network and loss function. Li, Cen et al. (2021) developed a meta-learning-based method that can fuse source images of different resolutions and produce a high-resolution fused image. Additionally, it utilizes a multi-task loss function to bolster feature learning of the fused image. Long et al. (2021) solved the image fusion task as a structure and intensity proportional maintenance problem and adopted two loss functions to enhance the feature extraction and fusion. Tang, Xiang et al. (2023) presented a darkness-free IVIF method to produce high-quality fused images with realistic color and contrast in night scenes. Ma et al. (2022) proposed a unified image fusion framework, termed SwinFusion. It integrates complementary and global information through attention-guided cross-domain modules while utilizing self-attention mechanisms and cross-domain attention mechanisms to extract specific and complementary features.

2.2.3. *GAN-based image fusion methods*

GAN-based methods leverage generative adversarial networks to perform unsupervised image fusion. This network comprises a generator and a discriminator. The generator aims to produce fused images from latent features, while the discriminator strives to distinguish fused images from source images. The generator produces images that match the source image distribution through adversarial training with the discriminator. FusionGAN (Ma, Yu et al., 2019) improves the fused image's texture by establishing a generative adversarial framework between the fused and visible images. Zhang, Yuan et al. (2021) preserved the contrast of thermal targets and the texture of source images by using a full-scale skip-connected generator, two Markovian discriminators, and a joint gradient loss. Rao et al. (2023) extracted compact and robust features from multimodal images in various adverse conditions and learned an adaptive equilibrium point for fusion with a quality assessment module.

2.3. *Task-driven image fusion methods*

Most existing fusion methods unilaterally focus on the visual quality but ignore the semantic information of the fused image (Tang, Zhang et al., 2023). To address this problem, Tang, Yuan, Ma (2022) cascaded the fusion network with the semantic segmentation network to supervise the fusion network so as to focus on the semantic information in the image. Sun et al. (2022) proposed a detection-driven image fusion network termed DetFusion. This network utilizes the target detection network to guide multimodal image fusion. Tang, Deng et al. (2022) proposed an image registration and fusion network termed SuperFusion. This model introduced a semantic segmentation network that prompted the network to uprate other high-level visual tasks. Tang, Zhang et al. (2023) embed a semantic segmentation network into the image fusion network to progressively inject semantic information into the fusion process. Wang et al. (2023) proposed an interactive enhancement paradigm for the joint infrared and visible image fusion and salient object detection tasks.

3. Proposed method

3.1. *Overview*

The structure of the SFINet for IVIF is shown in Fig. 3. The process covers four modules: (1) *multi-scale feature extraction*, (2) *dual attention feature fusion*, (3) *semantic feature interaction*, and (4) *image reconstruction*.

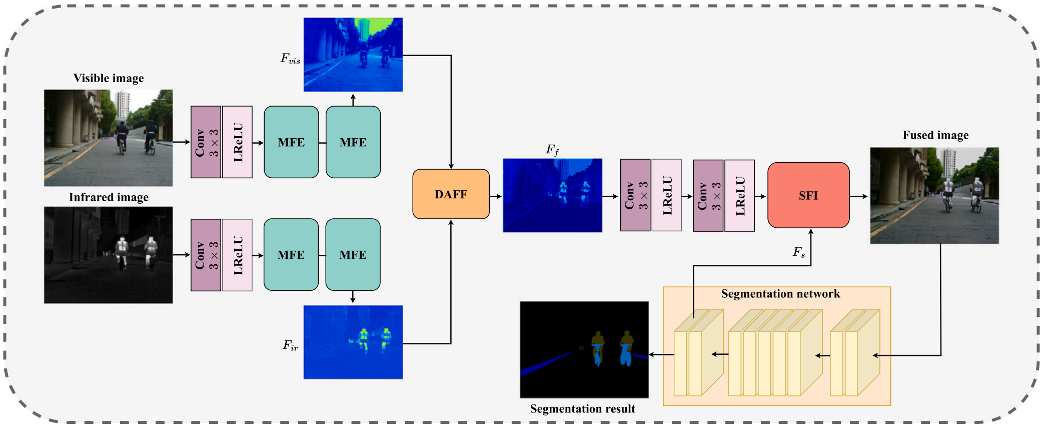


Fig. 3. Architecture of the proposed SFINet for infrared and visible image fusion.

Given a pair of infrared image $I_{ir} \in \mathbb{R}^{1\times H\times W}$ and visible image $I_{vis} \in \mathbb{R}^{3\times H\times W}$, we designed a MFE module to extract the complementary information from the two modal images. The process is denoted as,

$$\left\{F_{ir}, F_{vis}\right\} = \left\{\mathrm{MFE}(I_{ir}), \mathrm{MFE}(I_{vis})\right\}, \tag{1}$$

where F_{ir} and F_{vis} are features of infrared and visible images. The MFE module employs a densely connected structure, and it is composed of dilated convolutions with different dilation rates. Therefore, it can extract low-frequency global information from the source image. The residual branch uses a gradient operator and convolutional layers to extract high-frequency local information from the image.

Furthermore, the F_{ir} and the F_{vis} are fused by a DAFF module. It consists of spatial attention and self-attention units, which are used to explore the correlation between features and establish a close connection. The fused feature F_f is formulated as,

$$F_f = \mathrm{DAFF}(F_{ir}, F_{vis}). \tag{2}$$

Finally, to enhance the semantic information of the features and reconstruct them into a fused image, the F_f and the output features of the segmentation network (BANet) (Peng et al., 2021) interact through the SFI module. Moreover, during testing and the first round of iterative training ($i = 1$), the visible image is used as the input of the segmentation network. This process is formulated as,

$$\begin{cases} I_f^i = \mathrm{SFI}\left(F_f^i, N_s\left(I_{vis}\right)\right), i = 1 \\ I_f^i = \mathrm{SFI}\left(F_f^i, N_s\left(I_f^{i-1}\right)\right), i > 1, \end{cases} \tag{3}$$

where i is the number of iterations during training. N_s represents the segmentation network. I_f denotes the fused image.

3.2. Network architecture

3.2.1. Multi-scale feature extraction module

To extract multi-scale features of the source image, Li et al. (2020) designed a nest connection architecture network (NestFuse). This network employs a nest connection encoder model to extract multi-scale features. However, the nest connection encoder lacks attention to detailed features, which affects the balance between local details and the global context. By contrast, this paper constructs two MFE modules to process the infrared and visible images, respectively. This module employs densely connected dilated convolutions with varying dilation rates to extract multi-scale and low-frequency global features. Simultaneously, it enhances high-frequency local information through the residual gradient branch.

As shown in Fig. 4, the MFE module combines residual learning and dense connection structures. Additionally, a gradient operator is employed to extract detailed information from the images. The MFE module consists of two main branches, namely the main branch and the residual branch. The main branch contains five 3 × 3 dilated convolution units, with dilation rates $r = (1, 2, 3, 5, 7)$, and a 1 × 1 convolution layer. The residual branch comprises a gradient operator and a 1 × 1 convolution layer. The input feature F_{MFE}^{in} is fed to these two branches separately. The main branch extracts the multi-scale local and global features, and the residual branch extracts the edge details. Finally, the outputs of the main and residual branches are combined to output feature F_{MFE}^{out}.

3.2.2. Dual attention feature fusion module

The aforementioned NestFuse (Li et al., 2020) comprises spatial attention (SA) and channel attention (CA) modules to integrate the complementary features of source images. However, the features in the two attention modules are processed separately. This may lead to detail loss and information imbalance. In this work, we proposed the DAFF module by replacing the CA module to channel self-attention (CSA) module to focus on the global features and integrating the SA and CSA branches to fuse the feature adaptively. The DAFF module can learn the correlations between different modal features synchronously.

The DAFF module is illustrated in Fig. 5. The input of the DAFF module is the infrared feature F_{ir} and the visible feature F_{vis}. To achieve feature fusion, element-wise multiplication and element-wise addition are performed on the two features, respectively. Element-wise multiplication can extract the common information of the F_{ir} and F_{vis}, while element-wise addition can fuse the whole information of the F_{ir} and F_{vis}. Then, the spatial attention is applied to the element-wise addition features F_{sum} to generate feature F_{sa} that highlights salient objects and detailed information. Finally, we input the product of F_{sa} and the element-wise multiplication features F_{mul} elements into channel self-attention to enhance the features and generate the output of the DAFF module. It enables the DAFF module to adjust the weights of each channel adaptively. In summary, the process of the DAFF module is formulated as follows:

$$F_f = \mathrm{CSA}\big(\mathrm{SA}(F_{ir} \otimes F_{vis}) \otimes \mathrm{CBR}(F_{ir} \oplus F_{vis})\big), \tag{4}$$

where $\mathrm{SA}(\cdot)$ and $\mathrm{CSA}(\cdot)$ denote the spatial attention and channel self-attention, respectively. CBR is the operation composed by Convolutional layer, BatchNorm layer, and ReLU activation. $\otimes$ and $\oplus$ are element-wise multiplication and element-wise summation, respectively.

3.2.3. Semantic feature interaction module

In the PSFusion (Tang, Zhang et al., 2023), the semantic information is injected into the fusion network by a progressive semantic injection module (PSIM). To focus on the crucial features of the fused and semantic features, we proposed the SFI module by incorporating sequential CA and SA modules. The architecture of the SFI module is shown in

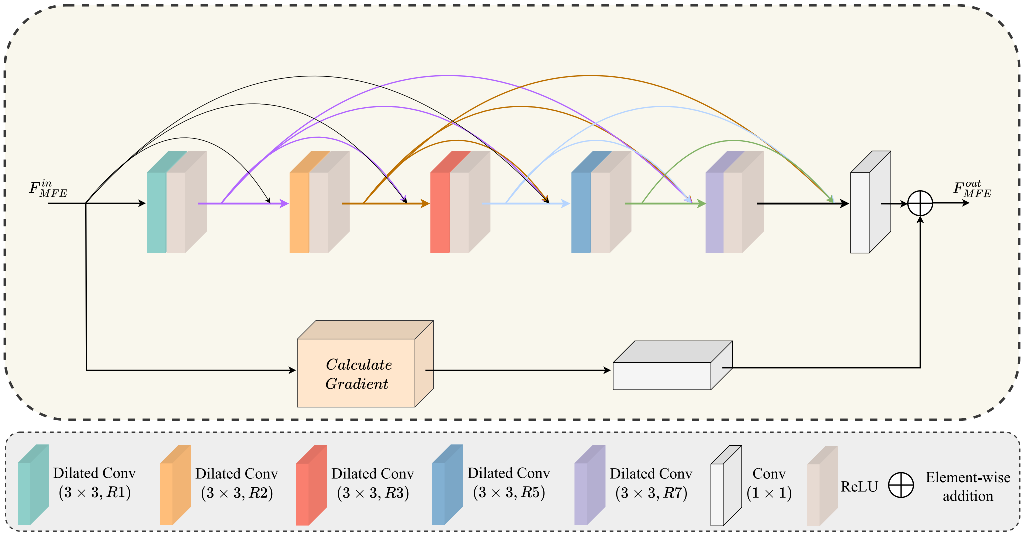


Fig. 4. Architecture of the MFE module.

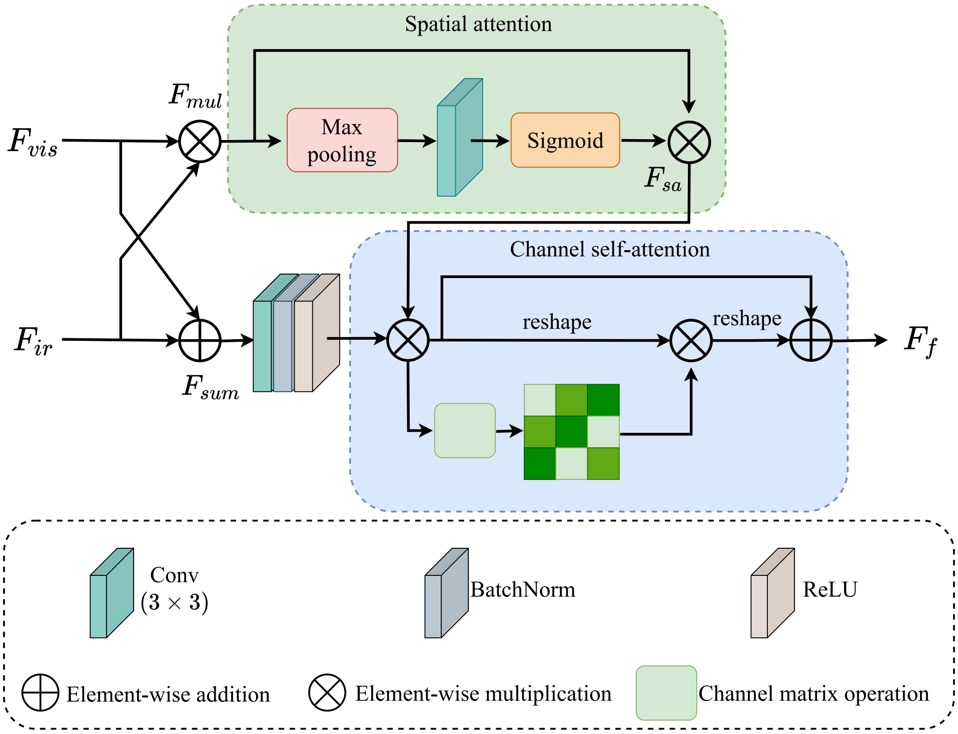


Fig. 5. Architecture of the DAFF module.

Fig. 6. The fusion features F_f and the semantic features F_s are first concatenated and then fed into a 1 × 1 convolution layer to generate the interacted features F_{inter}. Then, an element-wise multiplication is adopted to eliminate redundant information. Next, an element-wise addition operation is performed to aggregate the features. After that, a sequential CA and SA module are adopted to enhance the features. Finally, the residual connection is used to preserve the original information and generate the fused image I_f. The process of the SFI module is formulated as,

$$F_{inter} = \mathrm{Conv}_{1\times1}\left(\mathrm{Cat}\left(F_f, F_s\right)\right), \tag{5}$$

$$F_{inter2} = F_{inter} \otimes F_f + F_{inter} \otimes F_s, \tag{6}$$

$$I_f = Conv_{1\times1}\left(F_f + \mathrm{BN}\left(Conv_{3\times3}\left(\mathrm{SA}\left(\mathrm{CA}(F_{inter2})\right)\right)\right)\right), \tag{7}$$

where $Cat(\cdot)$ denotes the channel concatenation. ⊗ represents element-wise multiplication. $BN(\cdot)$ denotes a BatchNorm layer.

3.3. Loss function

The total loss function $\mathcal{L}_{\text{total}}$ consists of the image fusion loss $\mathcal{L}_f$ and the segmentation loss $\mathcal{L}_s$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_f + \alpha\mathcal{L}_s, \tag{8}$$

where α is the weight to dynamically regulate the relative significance of image fusion loss and segmentation loss, so as to avoid over-fitting. The parameter α gradually increases according to the joint adaptive training strategy of the low-level and high-level as the segmentation network adapts to the fusion model during the training process (Tang,

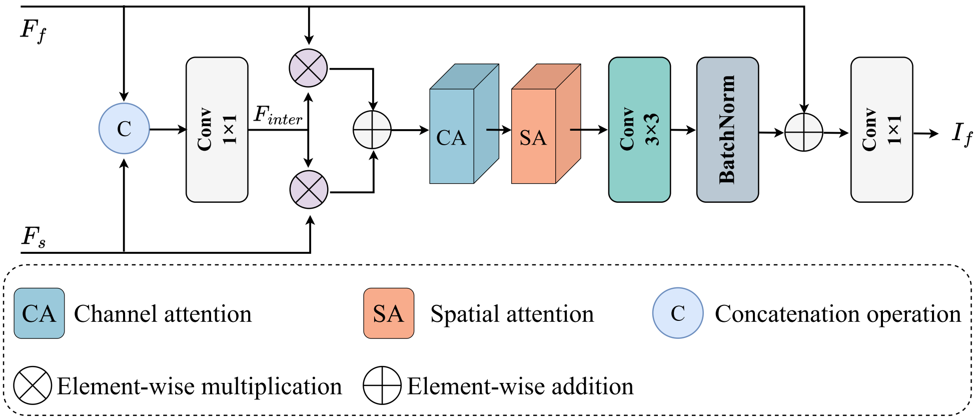


Fig. 6. Architecture of the SFI module.

Yuan, Ma, 2022). It is formulated as:

$$\alpha = \theta \times (m-1), \tag{9}$$

where m is the mth iteration. θ is a constant for the balance between semantic loss and content loss. Specifically, as the segmentation network increasingly fits the fusion model with more epochs, we progressively increase the semantic loss. This adjustment enables the semantic loss to guide the training of the fusion network more accurately as the training process.

The image fusion loss function comprises three loss functions, *i.e.*, pixel loss $\mathcal{L}_{pix}$, gradient loss $\mathcal{L}_{gra}$, and structural loss $\mathcal{L}_{ssim}$. The image fusion loss is formulated as,

$$\mathcal{L}_f = \lambda_1 \mathcal{L}_{pix} + \lambda_2 \mathcal{L}_{gra} + \lambda_3 \mathcal{L}_{ssim}, \tag{10}$$

where λ_1, λ_2, and λ_3 are dynamic factors that adjust the weights of the fusion loss function.

The pixel loss function aims to preserve as much information as possible from the source images in the fused image while minimizing distortion and artifacts. It is formulated as,

$$\mathcal{L}_{\text{pix}} = \frac{1}{HW}(\left\|I_f - I_{ir}\right\|_F^2 + \left\|I_f - I_{vis}\right\|_F^2), \tag{11}$$

where H and W are the height and width of the image, respectively. $\|\cdot\|_F$ denotes the Frobenius norm of the matrix.

The gradient loss enables the fused image to preserve the edge features of the input image and enhances the contrast and clarity of the fused image. It is defined as,

$$\mathcal{L}_{\text{gra}} = \|\nabla I_f - \max\left\{\nabla I_{ir}, \nabla I_{vis}\right\}\|_2, \tag{12}$$

where $\|\cdot\|_2$ stands for the ℓ_2-norm of the matrix. ∇ is the gradient operation, and $max\{\cdot,\cdot\}$ is maximum operator.

The Structural Similarity Index (SSIM) (Wang et al., 2004) is a widely adopted metric for measuring image distortion. It compares the similarity of images based on luminance, contrast, and structure aspects. To retain the essential features of source images, we constructed a structural loss function employing the SSIM. This ensures that the fused image is structurally consistent with the source images. It is formulated as,

$$\mathcal{L}_{\text{ssim}} = 1 - \text{SSIM}\left(I_f, \max\left\{I_{ir}, I_{vis}\right\}\right), \tag{13}$$

where SSIM$(\cdot)$ represents structural similarity index.

The segmentation loss function is composed of the main segmentation loss $\mathcal{L}_{main}$ and the auxiliary segmentation loss $\mathcal{L}_{aux}$. These two losses are expressed as,

$$\mathcal{L}_{main} = -\sum_{class} G\log(S), \tag{14}$$

$$\mathcal{L}_{aux} = -\sum_{class} G\log(\hat{S}), \tag{15}$$

where S denotes the predicted main segmentation label. $\hat{S}$ is the predicted auxiliary segmentation label. G represents the ground truth label. The segmentation loss function is formulated as follows,

$$\mathcal{L}_{semantic} = \mathcal{L}_{main} + \beta\mathcal{L}_{aux}, \tag{16}$$

where β is a loss weight that balances the main and auxiliary segmentation loss.

4. Experimental analysis

4.1. Benchmarks and implementation details

Three public datasets, namely MSRS (Tang, Yuan, Zhang et al., 2022), M3FD (Liu et al., 2022), and TNO (Toet, 2017) are adopted for qualitative and quantitative evaluation. The training set consists of 1083 pairs of images from the MSRS dataset. Additionally, we collect 361 pairs, 300 pairs, and 40 pairs of images from these three datasets as test sets. These images are normalized to $[0, 1]$. The image segmentation and fusion networks are conducted on the MSRS dataset. The MSRS dataset contains semantic labels for nine types of objects, including background, car, person, bike, curve, car stop, guardrail, color cone, and bump. The fusion network is trained with the initial learning rate of 0.001, a batch size of 4 samples. The Adam is adopted to optimize the fusion network under the guidance of the total loss. The initial learning rate for the segmentation network is set to 0.01. The method for updating the learning rate is to multiply the initial learning rate by a factor $(1 - \frac{epoch}{max_{epoch}})^{\gamma}$. The power γ of this factor is set to 0.9. The weight factors β, θ, λ_1, λ_2, and λ_3 are specified as 0.75, 1, 1, 100, and 10, respectively. The experiments are performed on an NVIDIA GeForce RTX 3090 Ti GPU with the PyTorch framework.

To determine the maximum number of iterations i_{max}, the convergence curves of image fusion loss and segmentation loss are drawn and depicted in Fig. 7. One can see that the loss function gradually decreases along with the increases of i. It indicates that the network is continuously optimizing and achieving better fusion results. Meanwhile, it shows that as the number of iterations i reached a certain value, the model parameters converged. In this work, we set the value of i_{max} to 4 by analyzing the trend of the loss function to ensure the model learns the optimal parameters.

4.2. Evaluation metrics

To evaluate the quality of the fused images, six quantitative indicators are employed, *i.e.,* Entropy (EN), Spatial Frequency (SF), Standard Deviation (SD), Average Gradient (AG), Overall Cross Entropy (OCE), and Edge Intensity (EI). EN indicates the amount of information and uncertainty in the fused image. SF reflects the amount of detail and

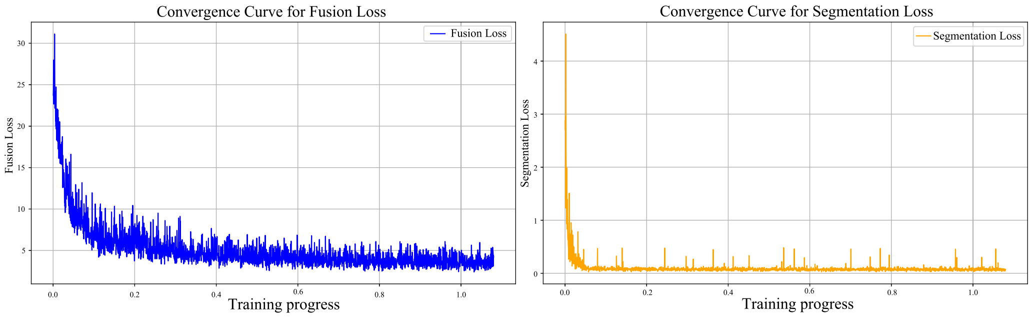


Fig. 7. Convergence curves of image fusion loss and segmentation loss.

contrasts in the fused image. SD represents the dispersion and variation of the pixel values in the fused image. AG denotes the rate of change and sharpness of the pixel values in the fused image. OCE is used to evaluate the amount of information retained from the original images. EI is to assess the clarity and sharpness of edges in the fused image. Higher values of EN, SF, SD, AG, OCE, and EI indicate better performance.

Additionally, Intersection over Union (IoU) and mean Intersection over Union (mIoU) are utilized to assess the contribution of the fused images to semantic segmentation tasks. IoU measures the overlap between the predicted segmentation and the ground truth. mIoU provides an average measure over multiple classes to comprehensively evaluate segmentation accuracy. Higher values of IoU and mIoU indicate more accuracy in semantic segmentation tasks.

4.3. Performance analysis

To assess the effectiveness of the proposed SFINet, we conducted comparative experiments with nine state-of-the-art (SOTA) methods. The competitors include CDDFuse (Zhao, Bai et al., 2023), NestFuse (Li et al., 2020), PSFusion (Tang, Zhang et al., 2023), RFN-Nest (Li, Wu et al., 2021), SeAFusion (Tang, Yuan, Ma, 2022), SuperFusion (Tang, Deng et al., 2022), UMF-CMGR (Di et al., 2022), and YDTR (Tang, He et al., 2022).

4.3.1. Qualitative evaluation

To evaluate the effectiveness of SFINet in seamlessly integrating full-time images and concurrently merging complementary information to enhance visual quality, we selected a daytime case and a nighttime case for qualitative evaluation. The qualitative evaluation results are illustrated in Fig. 8. In the case of daytime, the VIS image contains more information, and infrared images can complement the significant information in VIS images. Therefore, an excellent fusion method should incorporate rich detailed information from VIS images and the prominent targets from IR images. Also, it is observed that RFN-Nest, UMF-CMGR, and YDTR perform poorly in preserving edge details. Meanwhile, the salient information of RFN-Nest is weakened. PSFusion is severely disturbed by irrelevant information from the IR image. In contrast, the proposed SFINet can effectively preserve both the detailed information of the VIS image and the salient information of the IR image.

In the nighttime scene, the reduction of visual information poses a challenge. Fig. 8 shows that the proposed model not only preserves the salient information inside the red box in the image but also displays the target information of persons in the distance. This is attributed to the proposed MFE module. It can effectively preserve the multi-scale information in the source image. Furthermore, we enlarged a dark area with a green box, where the proposed method can clearly show the details of a car. This can be attributable to the fact that the DFF module can adaptively integrate complementary information from the source images. Meanwhile, the SFI module also contributes rich semantic information.

4.3.2. Quantitative evaluation

The quantitative results on the MSRS dataset are shown in Table 1. It shows that the proposed SFINet ranks first in EN and SD. It indicates that SFINet excels in gradient information preservation, contrast, and effective information retention. Meanwhile, the SFINet ranks second only behind PSFusion in SF, AG, and EI and ranks third-best in OCE. It proves that the proposed SFINet can produce fusion images with excellent visual effects and effectively retain valuable information from the source images.

4.4. Generalization evaluation

4.4.1. Generalization evaluation on the M3FD dataset

To evaluate the generalization ability of the SFINet, we conducted experiments on the M3FD dataset. The M3FD dataset consists of images captured in various scenarios. The images have diverse levels of brightness, contrast, and noise. Therefore, an effective fusion approach necessitates adaptability to different conditions and preserves the complementary information from both source images. Fig. 9 displays the qualitative results on the M3FD dataset. It shows that only SFINet can retain good visual contrast in the fused images, and other methods cannot eliminate the blurring issue in visible images. On the other hand, CDDFuse, RFN-Nest, PSFusion, SeAFusion, SuperFusion, SwinFusion, UMF-CMGR, and YDTR cannot clearly retain the structure of the buildings. In contrast, the SFINet can effectively utilize the complementary information of the source image to generate a fused image with high contrast and rich details. This can benefit the proposed three modules, *i.e.*, MFE, DAFF, and SFI modules, that are responsible for extracting multi-scale features, learning the correlation between source images, and enhancing semantic information.

The quantitative results of the proposed method with nine SOTA methods on 300 pairs of images from the M3FD dataset are displayed in Table 2. It indicates that the proposed SFINet achieved superior performance in terms of SD. This demonstrates that the SFINet effectively utilizes the complementary information in the source images, resulting in less distortion during the fusion process. Moreover, the SFINet is generally inferior to PSFusion in EN, SF, AG, and EI, while maintaining an average level in OCE. There are two reasons for the result. On the one hand, the proposed method eliminates redundant information in the source images during the fusion process, thereby leading to some inconsistencies between the fused image and the source images. On the other hand, compared to SFINet, PSFusion has more parameters to enhance the network's generalization ability.

4.4.2. Generalization evaluation on the TNO dataset

The qualitative comparison of different methods on the TNO dataset is presented in Fig. 10. The red boxes highlight significant shortcomings in preserving salient targets for NestFuse, RFN-Nest, SuperFusion, UMF-CMGR, and YDTR. Furthermore, other image fusion methods lead to substantial spectral interference in background segments and erase intricate background information. In contrast, SFINet can effectively

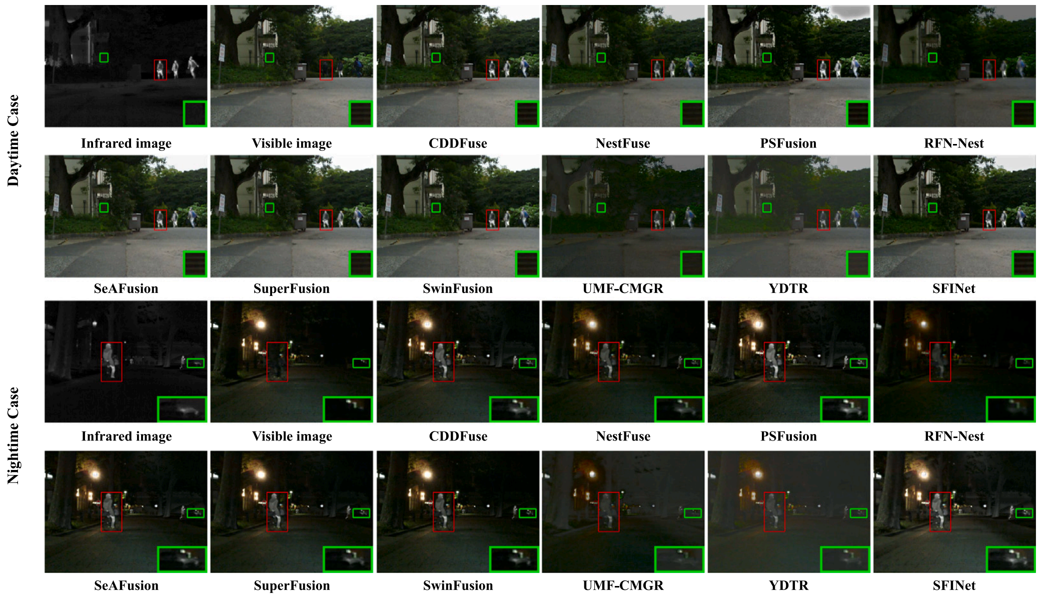


Fig. 8. Qualitative comparison of the proposed SFINet with different methods on the MSRS dataset.

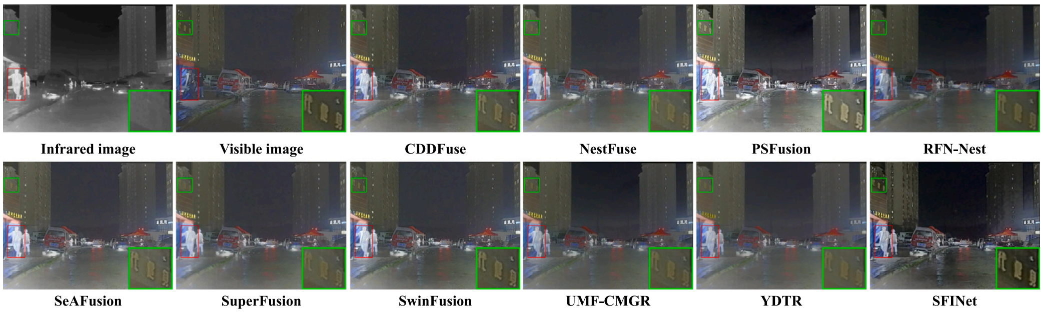


Fig. 9. Qualitative comparison of the proposed SFINet with different methods on the M3FD dataset.

Table 1
Quantitative results of the proposed method with nine SOTA methods on 361 pairs of images from the MSRS dataset. The best, second-best, and third-best methods are marked in red, blue, and green, respectively.

Methods	EN ↑	SF ↑	SD ↑	AG ↑	OCE ↑	EI ↑
CDDFuse (Zhao, Bai et al., 2023)	6.699	0.045	8.436	3.744	0.990	39.808
NestFuse (Li et al., 2020)	6.501	0.038	8.217	3.118	1.015	33.210
PSFusion (Tang, Zhang et al., 2023)	6.779	0.052	8.385	4.446	0.908	47.082
RFN-Nest (Li, Wu et al., 2021)	6.175	0.024	7.786	2.143	1.009	23.285
SeAFusion (Tang, Yuan, Ma, 2022)	6.652	0.044	8.377	3.697	1.148	39.551
SuperFusion (Tang, Deng et al., 2022)	6.587	0.042	8.335	3.394	1.041	36.227
SwinFusion (Ma et al., 2022)	6.619	0.043	8.409	3.546	1.052	37.770
UMF-CMGR (Di et al., 2022)	5.600	0.028	6.181	2.161	3.380	22.454
YDTR (Tang, He et al., 2022)	5.645	0.029	6.828	2.201	1.140	23.165
SFINet (Ours)	6.785	0.045	8.824	3.827	1.440	40.914

Table 2
Quantitative results of the proposed method with nine SOTA methods on 300 pairs of images from the M3FD dataset. The best, second-best, and third-best methods are marked in red, blue, and green, respectively.

Methods	EN ↑	SF ↑	SD ↑	AG ↑	OCE ↑	EI ↑
CDDFuse (Zhao, Bai et al., 2023)	6.904	0.058	9.972	4.863	1.644	50.141
NestFuse (Li et al., 2020)	6.804	0.044	9.608	3.751	1.805	38.857
PSFusion (Tang, Zhang et al., 2023)	7.399	0.081	9.776	6.917	1.487	71.355
RFN-Nest (Li, Wu et al., 2021)	6.862	0.030	9.230	2.870	1.651	30.686
SeAFusion (Tang, Yuan, Ma, 2022)	6.846	0.055	9.863	4.782	1.743	49.684
SuperFusion (Tang, Deng et al., 2022)	6.692	0.044	9.846	3.782	1.538	38.863
SwinFusion (Ma et al., 2022)	6.799	0.053	9.858	4.599	1.655	47.564
UMF-CMGR (Di et al., 2022)	6.702	0.034	9.146	2.944	1.751	30.469
YDTR (Tang, He et al., 2022)	6.547	0.040	9.194	3.306	1.589	34.130
SFINet (Ours)	6.997	0.069	9.975	6.138	1.541	63.924

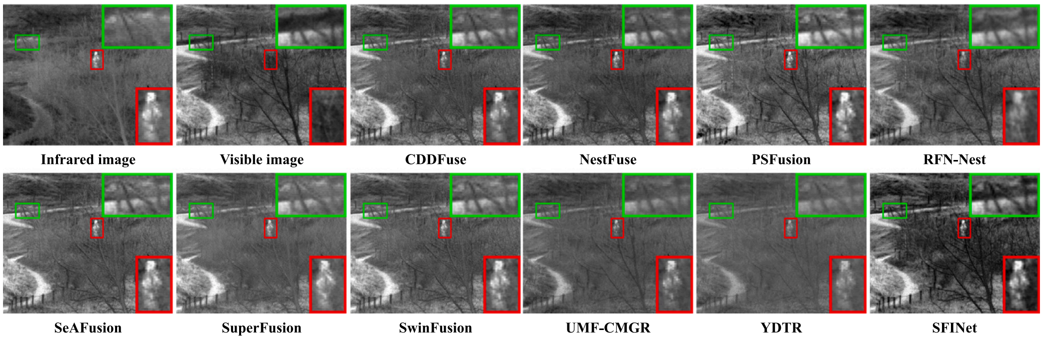


Fig. 10. Qualitative comparison of the proposed SFINet with different methods on the TNO dataset.

Table 3
Quantitative results of the proposed method with nine SOTA methods on 40 pairs of images from the TNO dataset. The best, second-best, and third-best models are marked in red, blue, and green, respectively.

Methods	EN ↑	SF ↑	SD ↑	AG ↑	OCE ↑	EI ↑
CDDFuse (Zhao, Bai et al., 2023)	7.077	0.049	9.395	4.715	1.698	45.721
NestFuse (Li et al., 2020)	7.011	0.039	9.364	3.835	1.604	37.910
PSFusion (Tang, Zhang et al., 2023)	7.326	0.052	9.596	5.377	1.078	54.092
RFN-Nest (Li, Wu et al., 2021)	6.991	0.023	9.379	2.682	1.701	28.644
SeAFusion (Tang, Yuan, Ma, 2022)	7.136	0.048	9.562	5.011	1.507	50.495
SuperFusion (Tang, Deng et al., 2022)	6.754	0.036	9.095	3.582	1.278	34.691
SwinFusion (Ma et al., 2022)	6.899	0.042	9.356	4.205	1.294	41.250
UMF-CMGR (Di et al., 2022)	6.559	0.032	8.723	2.988	1.445	29.372
YDTR (Tang, He et al., 2022)	6.432	0.030	8.829	2.788	1.076	27.297
SFINet (Ours)	7.251	0.051	9.758	5.443	1.640	55.509

retain the texture details of the visible image and successfully maintain the clarity of salient targets.

The quantitative comparison results on the TNO dataset are displayed in Table 3. It shows that the proposed model ranks first in SD, AG, and EI metrics. This achievement demonstrates that the SFINet can present rich texture details and salient object information in the fusion images while ensuring the visual quality of the fused image. In EN and SF, the proposed method ranks second only behind PSFusion. In addition, the SFINet is inferior to RFN-Nest and CDDFuse, but the overall level is still above average.

4.5. Segmentation comparison and analysis

In this section, we constructed comprehensive experiments to verify the contribution of the SFINet to the semantic segmentation task. In particular, the experiments are conducted on image-level and feature-level fusion segmentation tasks on the MSRS dataset.

4.5.1. Image-level fusion segmentation comparison

We utilized the pre-trained Segformer (Xie et al., 2021) as the segmentation model to test the contribution of the SFINet and image-level fusion competitors to the semantic segmentation task. Fig. 11 shows the qualitative results of the segmentation task in two cases. In the daytime case, Segformer fails to capture the content of the "person" on the left in the results of CDDFuse, PSFusion, and RFN-Nest, and the segmentation of the "person" was inaccurate in the results of SwinFusion and UMF-CMGR. In the nighttime case, except for the visible image and the results of RFN-Nest, SeAFusion, and SFINet, other methods suffer from the semantic content loss of "curve". The results of the CDDFuse, NestFuse, PSFusion, SuperFusion and SwinFusion also exhibited the issue of losing "person" semantic content. The SFINet can accurately segment the objects of interest, *e.g.,* cars, persons, and curves, in both daytime and nighttime.

The quantitative semantic segmentation results on the MSRS dataset are shown in Table 4. The results show that the SFINet ranks first in classes of "Car", "Person", "Bike", "Car stop", "Color cone", and second best in "Unlabelled", and "Curve" in terms of IoU. Meanwhile, the SFINet performs best in terms of mIoU. These findings demonstrate the effectiveness of fused images in enhancing segmentation performance.

4.5.2. Feature-level fusion segmentation comparison

To comprehensively verify the contribution of the proposed method to the semantic segmentation task, we compared the SFINet with six SOTA feature-level fusion segmentation methods using the trained BANet (Peng et al., 2021) by SFINet as the segmentation network. The competitors include MFNet (Ha et al., 2017), GMNet (Zhou et al., 2021), EGFNet (Zhou et al., 2022), MDRNet (Zhao, Liu et al., 2023), LASNet (Li et al., 2022) and RTFNet (Sun et al., 2019). Fig. 12 shows the visualization of semantic segmentation results on the MSRS dataset. The proposed method can continuously achieve high-precision classification of objects in different scenes. By contrast, the other models fail to classify the obstacles and cars in daytime scenes and the person in nighttime scenes.

Table 5 displays the qualitative semantic segmentation results of different competitors on the MSRS dataset. The results prove that SFINet outperforms the competitors in IoU across all categories and mIoU. Specifically, compared to the second-best model, LASNet, the mIoU is improved by 29.93% in mIoU. It proves that the proposed method can effectively maintain and enhance visual quality while fully utilizing the information from different modality images.

4.6. Computational complexity analysis

To verify the efficiency of the proposed method, we examined the running times and the parameters of all methods on MSRS, M3FD, and TNO datasets. As shown in Table 6, the proposed SFINet has fewer parameters than the PSFusion. Meanwhile, the running time is competitive compared with other SOTA models.

4.7. Ablation studies

Ablation studies were conducted on the MSRS dataset to assess the effectiveness of each module in SFINet. The parameters and dataset settings are consistent for all the ablation experiments.

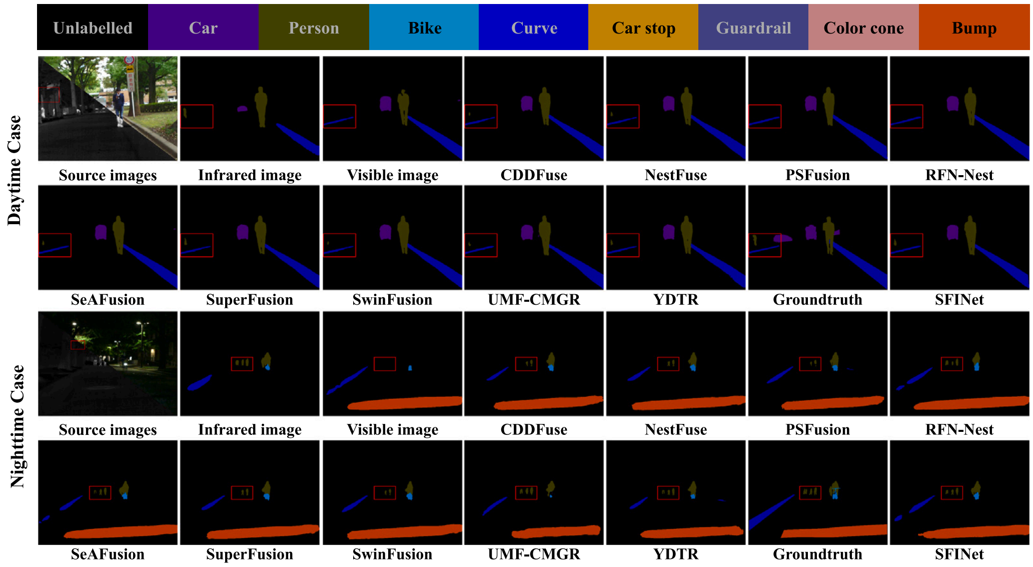


Fig. 11. Qualitative semantic segmentation comparisons of the SFINet and nine competitors in two cases on the MSRS dataset.

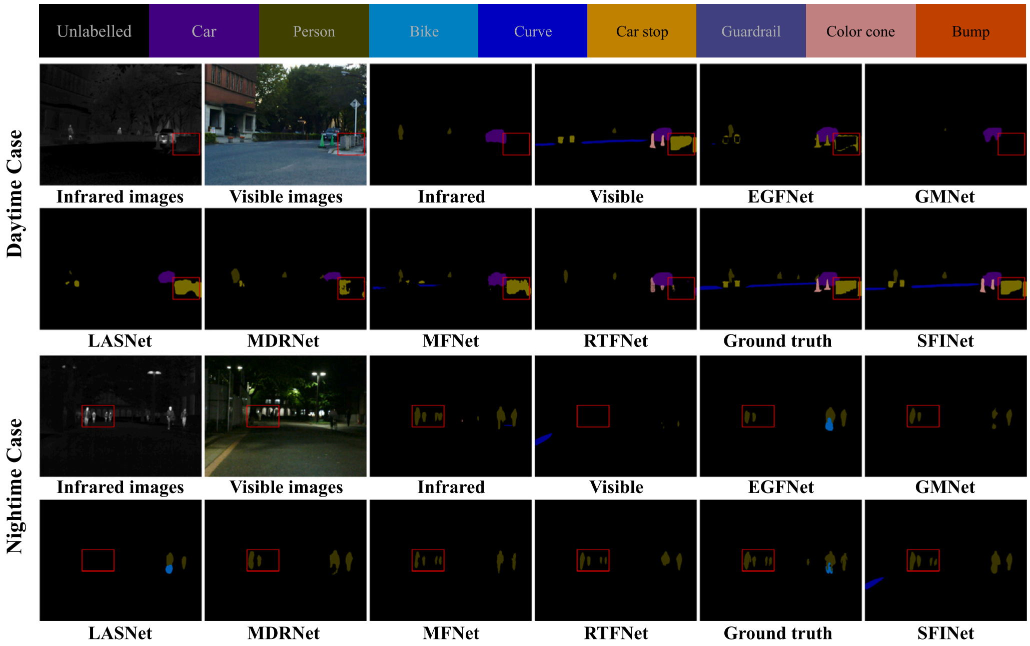


Fig. 12. Qualitative semantic segmentation comparisons of the SFINet and six feature-level fusion segmentation methods in two cases on the MSRS dataset.

Table 4
Quantitative semantic segmentation results of various methods on the MSRS dataset. The segmentation model utilizes the Segformer pre-trained on the MSRS dataset. The best and second-best methods are marked in red and blue, respectively.

Methods	IoU ↑									mIoU ↑
	Unlabelled	Car	Person	Bike	Curve	Car stop	Guardrail	Color cone	Bump	
Infrared	96.00	56.76	70.62	32.32	34.93	22.40	0.	15.09	37.10	40.58
Visible	97.91	87.50	38.14	70.49	51.72	73.16	86.46	64.18	79.28	72.10
CDDFuse (Zhao, Bai et al., 2023)	98.51	89.32	72.44	72.01	60.18	77.20	87.59	64.17	79.77	77.91
NestFuse (Li et al., 2020)	98.53	89.98	73.21	72.61	59.96	75.64	86.66	64.33	80.17	77.90
PSFusion (Tang, Zhang et al., 2023)	98.54	89.55	73.81	72.86	60.59	77.55	85.59	65.00	80.19	78.22
RFN-Nest (Li, Wu et al., 2021)	98.64	89.96	72.44	72.78	61.53	77.83	85.42	62.54	79.38	77.82
SeAFusion (Tang, Yuan, Ma, 2022)	98.53	89.58	73.05	72.73	59.96	76.83	87.00	64.62	80.01	78.03
SuperFusion (Tang, Deng et al., 2022)	98.52	89.74	73.29	72.78	59.53	75.38	86.96	64.08	79.64	77.77
SwinFusion (Ma et al., 2022)	98.50	89.60	72.26	72.18	59.56	76.31	87.40	64.36	79.95	77.79
UMF-CMGR (Di et al., 2022)	98.28	87.79	71.61	68.72	50.10	74.62	76.65	60.13	74.52	73.60
YDTR (Tang, He et al., 2022)	98.36	88.65	71.89	71.51	51.77	71.59	85.68	59.94	77.17	75.17
SFINet (Ours)	98.60	90.28	75.19	73.10	61.49	77.89	85.51	65.82	79.73	78.62

Table 5
Quantitative semantic segmentation results of the proposed method and six SOTA multimodal segmentation methods on the MSRS dataset. Red represents the best results, and blue represents the second-best results.

Methods	IoU ↑									mIoU ↑
	Unlabelled	Car	Person	Bike	Curve	Car stop	Guardrail	Color cone	Bump	
MFNet (Ha et al., 2017)	96.56	70.05	54.39	47.17	21.74	37.09	40.78	22.91	24.79	46.16
GMNet (Zhou et al., 2021)	96.30	70.47	58.93	33.17	6.98	36.38	20.68	18.33	41.22	42.5
EGFNet (Zhou et al., 2022)	97.69	84.91	61.78	60.22	39.67	0.59	0.	1.28	1.33	38.61
MDRNet (Zhao, Liu et al., 2023)	94.42	49.55	58.52	41.1	31.7	38.26	20.42	38.17	19.48	43.51
LASNet (Li et al., 2022)	96.94	80.59	47.97	51.32	42.96	51.07	33.91	47.74	53.91	56.27
RTFNet (Sun et al., 2019)	97.35	83.53	71.91	49.82	33.12	48.41	0.	26.59	23.58	48.26
SFINet (Ours)	98.37	89.18	72.01	69.55	58.16	70.98	67.03	63.13	69.61	73.11

Table 6
Parameter and running times comparisons of all methods on the MSRS, M3FD, and TNO datasets.

Methods	Parameter (M)	Time (s)		
		MSRS	M3FD	TNO
CDDFuse (Zhao, Bai et al., 2023)	1.787	0.566	1.409	1.041
NestFuse (Li et al., 2020)	2.733	0.518	1.050	0.435
PSFusion (Tang, Zhang et al., 2023)	45.909	0.177	0.324	0.154
RFN-Nest (Li, Wu et al., 2021)	7.524	0.176	0.308	0.164
SeAFusion (Tang, Yuan, Ma, 2022)	13.061	0.147	0.241	0.112
SuperFusion (Tang, Deng et al., 2022)	1.962	0.110	0.332	0.107
SwinFusion (Ma et al., 2022)	0.974	1.000	2.294	0.924
UMF-CMGR (Di et al., 2022)	0.629	0.221	0.322	0.348
YDTR (Tang, He et al., 2022)	0.218	0.215	0.450	0.173
SFINet (Ours)	13.107	0.172	0.314	0.143

Table 7
Quantitative ablation results of the proposed model on the MSRS dataset. The optimal results are marked in red.

Models	MFE	DAFF	SFI	EN ↑	SF ↑	SD ↑	AG ↑	OCE ↑	EI ↑
M1	✗	✓	✓	6.7229	0.0441	8.4147	3.6503	2.2828	39.0884
M2	✓	✓	✗	6.7248	0.0445	8.4826	3.8005	1.8870	40.6333
M3	✓	✗	✓	6.6449	0.0443	8.3770	3.7467	1.1695	39.9445
M4	✓	✗	✓	6.6505	0.0442	8.3772	3.7380	1.1751	39.8157
M5	✓	✗	✓	6.6313	0.0439	8.3687	3.6583	1.1093	39.0801
SFINet	✓	✓	✓	6.7851	0.0453	8.8242	3.8279	1.4395	40.9143

4.7.1. Analysis of the specific modules

We replaced each module with convolutional layers of the same depth. The results of the quantitative ablation are displayed in Table 7. The results prove that replacing the MFE module (termed M1) led to a notable deterioration in EN, SF, SD, AG, and EI metrics. This verifies the critical role of the MFE module in extracting multi-scale features from the source image. Similarly, when the SFI module was replaced (termed M2), the values of EN, SF, SD, AG, and EI decreased. It proves the essential role of the SFI module in enhancing semantic information and the visual quality of the fused image. It is worth mentioning that the OCE indicator increases after the MFE module and SFI module are replaced. This is because convolutional layers excel at extracting local detail information, and increasing the number of convolutional layers enhances the feature representation ability of the model. The configurations that replace the DAFF module with three traditional fusion strategies, namely summation (termed M3), average (termed M4), and L1-norm (termed M5), also perform worse than the proposed SFINet. It proves that the DAFF module is practical for learning the correlation between the source images and achieving adaptive feature fusion.

4.7.2. Analysis of the segmentation network

To validate the contribution of the segmentation network to the fusion task, we conducted comprehensive experiments. These included removing the segmentation network, employing Segformer (Xie et al., 2021) as the segmentation network, and utilizing BANet (Peng et al., 2021) as the segmentation network.

Table 8 presents the quantitative comparison results employing the different segmentation models. When the semantic segmentation network is removed (w/o), the EN, SF, SD, AG, and EI metrics decline. This indicates that the semantic segmentation network plays a critical role in enhancing image fusion. By providing additional semantic information, it significantly improves the fused image in terms of detail retention and information preservation. By including the auxiliary segmentation models, our method demonstrated outstanding performance in EN and SD metrics. It can preserve image information and improve edge clarity. Among the two segmentation variants, BANet offers a well-balanced performance that enhances semantic information while retaining image details. Additionally, BANet has only 12.894M. which makes BANet more suitable for use in resource-constrained environments. It is worth noticing that our framework is scalable to use various segmentation models.

5. Conclusion and future work

This work proposes a semantic feature interactive learning network (SFINet) for full-time infrared and visible image fusion. The SFINet consists of an image fusion network and an image segmentation network. The image fusion network incorporates the MFE and DAFF modules to effectively extract and integrate multi-scale local and global information from source images. Additionally, an SFI module is built to interact with the semantic features from the image segmentation network and fused features from the image fusion network. Comparative experiments on three datasets demonstrate that the SFINet outperforms the SOTA methods subjectively and objectively. The evaluation experiments on segmentation performance further highlight the effectiveness of the SFINet in enhancing the performance of segmentation tasks. Although SFINet has significant performance improvements in infrared and visible image fusion, the complexity increases during the interaction between the fusion model and the segmentation model. It hinders

Table 8
Quantitative ablation results of the segmentation model on the MSRS dataset. "w/o" means removing the segmentation model. The optimal results are marked in red.

Segmentation model	EN ↑	SF ↑	SD ↑	AG ↑	OCE ↑	EI ↑	Time (s)	Parameter (M)
w/o	6.715	0.044	8.461	3.743	1.578	39.932	0.158	–
Segformer (Xie et al., 2021)	6.755	0.090	8.430	4.545	1.222	45.881	0.196	44.605
BANet (Peng et al., 2021)	6.785	0.045	8.824	3.827	1.440	40.914	0.172	12.894

the model from being applied to embedded systems or mobile devices. Thus, future research is expected to focus on model compression and optimization.

CRediT authorship contribution statement

Wenhao Song: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Qilei Li:** Investigation, Data curation. **Mingliang Gao:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. **Abdellah Chehri:** Writing – review & editing, Supervision. **Gwanggil Jeon:** Conceptualization, Resources, Data curation.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62101310).

References

Di, W., Jinyuan, L., Xin, F., & Liu, R. (2022). Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *International joint conference on artificial intelligence* (pp. 3508–3515).

Fu, Z., Wang, X., Xu, J., Zhou, N., & Zhao, Y. (2016). Infrared and visible images fusion based on RPCA and NSCT. *Infrared Physics & Technology*, *77*, 114–123.

Gan, W., Wu, X., Wu, W., Yang, X., Ren, C., He, X., & Liu, K. (2015). Infrared and visible image fusion with the use of multi-scale edge-preserving decomposition and guided image filter. *Infrared Physics & Technology*, *72*, 37–51.

Gao, M., Zhou, Y., Zhai, W., Zeng, S., & Li, Q. (2023). SaReGAN: a salient regional generative adversarial network for visible and infrared image fusion. *Multimedia Tools and Applications*, 1–13.

Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., & Harada, T. (2017). Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ international conference on intelligent robots and systems* (pp. 5108–5115). IEEE.

Jian, L., Yang, X., Liu, Z., Jeon, G., Gao, M., & Chisholm, D. (2020). SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–15.

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.

Li, H., Cen, Y., Liu, Y., Chen, X., & Yu, Z. (2021). 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.

Li, G., Wang, Y., Liu, Z., Zhang, X., & Zeng, D. (2022). RGB-T semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(3), 1223–1235.

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

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.

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.

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. 5802–5811).

Long, Y., Jia, H., Zhong, Y., Jiang, Y., & Jia, Y. (2021). RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Information Fusion*, *69*, 128–141.

Ma, J., Ma, Y., & Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, *45*, 153–178.

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, *9*, 1200–1217.

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.

Mitchell, H. B. (2010). *Image fusion: theories, techniques and applications*. Springer Science & Business Media.

Peng, C., Tian, T., Chen, C., Guo, X., & Ma, J. (2021). Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation. *Neural Networks*, *137*, 188–199.

Rao, Y., Wu, D., Han, M., Wang, T., Yang, Y., Lei, T., Zhou, C., Bai, H., & Xing, L. (2023). AT-GAN: A generative adversarial network with attention and transition for infrared and visible image fusion. *Information Fusion*, *92*, 336–349.

Song, W., Gao, M., Li, Q., Guo, X., Wang, Z., & Jeon, G. (2024). Optimizing nighttime infrared and visible image fusion for long-haul tactile internet. *IEEE Transactions on Consumer Electronics*, *70*, 4277–4286.

Song, W., Zhai, W., Gao, M., Li, Q., Chehri, A., & Jeon, G. (2024). Multiscale aggregation and illumination-aware attention network for infrared and visible image fusion. *Concurrency Computations: Practice and Experience*, *36*, Article e7712.

Sun, Y., Cao, B., Zhu, P., & Hu, Q. (2022). Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 4003–4011).

Sun, Y., Zuo, W., & Liu, M. (2019). Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, *4*(3), 2576–2583.

Tang, L., Deng, Y., Ma, Y., Huang, J., & Ma, J. (2022). SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, *9*(12), 2121–2137.

Tang, W., He, F., & Liu, Y. (2022). YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Transactions on Multimedia*.

Tang, L., Xiang, X., Zhang, H., Gong, M., & Ma, J. (2023). DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion*, *91*, 477–493.

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.

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.

Tang, L., Zhang, H., Xu, H., & Ma, J. (2023). Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, *99*, Article 101870.

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

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.

Wang, D., Liu, J., Liu, R., & Fan, X. (2023). An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, *98*, Article 101828.

Wu, M., Ma, Y., Fan, F., Mei, X., & Huang, J. (2020). Infrared and visible image fusion via joint convolutional sparse representation. *Journal of the Optical Society of America A*, *37*(7), 1105–1115.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

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.

Xu, H., Zhang, H., & Ma, J. (2021). Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, *7*, 824–836.

Yan, X., Qin, H., Li, J., Zhou, H., & Zong, J.-g. (2015). Infrared and visible image fusion with spectral graph wavelet transform. *Journal of the Optical Society of America A*, *32*(9), 1643–1652.

Zhang, Q., Liu, Y., Blum, R. S., Han, J., & Tao, D. (2018). Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*, *40*, 57–75.

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

Zhang, H., Yuan, J., Tian, X., & Ma, J. (2021). GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators. *IEEE Transactions on Computational Imaging*, *7*, 1134–1147.

Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., & Van Gool, L. (2023). Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5906–5916).

Zhao, S., Liu, Y., Jiao, Q., Zhang, Q., & Han, J. (2023). Mitigating modality discrepancies for RGB-T semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhou, W., Dong, S., Xu, C., & Qian, Y. (2022). Edge-aware guidance fusion network for rgb–thermal scene parsing. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3571–3579).

Zhou, W., Liu, J., Lei, J., Yu, L., & Hwang, J.-N. (2021). GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, *30*, 7790–7802.

Zhou, Z., Wang, B., Li, S., & Dong, M. (2016). Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Information Fusion*, *30*, 15–26.

Zou, D., & Yang, B. (2023). Infrared and low-light visible image fusion based on hybrid multiscale decomposition and adaptive light adjustment. *Optics and Lasers in Engineering*, *160*, Article 107268.