

Article

Significance-Preserving Progressive Network for Infrared and Visible Image Fusion

Jingsui Li , Xiaorun Li * , Shu Xiang  and Shuhan Chen 

College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China; lijingsui@zju.edu.cn (J.L.); 12210076@zju.edu.cn (S.X.); 11410057@zju.edu.cn (S.C.)

* Correspondence: lxrlly@zju.edu.cn

Highlights

What are the main findings?

- A significance-preserving progressive network (SiPFusion) is proposed, featuring a progressive extraction process that integrates hierarchical multi-scale local features and Transformer-based global dependencies.
- A novel significance-preserving fusion module with a spatial selection mechanism and a complementary significance similarity loss function were uniquely designed.

What are the implications of the main finding?

- The progressive feature extraction and spatial selection mechanism effectively coordinate heterogeneous global-local features, ensuring the adaptive integration of target saliency and texture details.
- The designed modules improve image fidelity and structural consistency in complex scenes, enabling SiPFusion to achieve strong overall performance across three benchmarks, with leading results on most metrics while maintaining competitive artifact suppression performance.

Abstract

Fusing infrared and visible images can effectively compensate for the inherent limitations of each modality in different scenes, resulting in fused images that contain richer information. However, existing methods often struggle to balance global dependency modeling with local detail preservation and to effectively coordinate heterogeneous local and global features during fusion. To address these issues, this paper proposes a Significance-Preserving Progressive Fusion Network (SiPFusion). First, a progressive feature extraction framework was designed, which hierarchically extracts multi-scale local features using CNNs and then models long-range dependencies across scales via a Transformer-based global module. To adaptively integrate local-global complementary features, a significance-preserving fusion module was designed to obtain significance attention maps with a spatial selection mechanism, enabling dynamic fusion of multi-source features. Furthermore, we propose a significance similarity loss function that leverages intermediate feature guidance to enhance structural consistency and preserve salient-region information in the fused image. Extensive experiments on the MSRS, RoadScene, and TNO datasets demonstrate that SiPFusion achieves competitive visual quality and strong overall quantitative performance against 15 state-of-the-art fusion methods, obtaining leading results on most evaluated metrics.

Keywords: infrared and visible image fusion; autoencoder network; Transformer; significance-preserving; attention map



Academic Editors: Bin Yang, Xin Ye and Jing Li

Received: 26 May 2026

Revised: 1 July 2026

Accepted: 10 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Infrared and visible image fusion is an important branch of image fusion and a research hotspot worldwide, playing a crucial role in both civilian and military applications [1–3]. Visible images are generated based on the reflection of light, with a wavelength range of approximately 380 nm to 780 nm aligning with human visual perception [4]. They provide detailed resolution and abundant texture details. However, they are highly vulnerable to challenging weather conditions, including poor visibility, nighttime [5], and fog. In contrast, infrared images, generated based on the radiation emitted by objects with a wavelength range of about 8 μm to 14 μm [6], can capture thermal targets in all weather conditions, due to their independence from external illumination sources. However, infrared images suffer from low resolution, high noise, poor contrast, and blurry edges, making it difficult to reveal texture details and providing poor visual quality. [7,8] By fusing these two types of images, the expressive power of the scene is greatly enhanced. The fused image not only shows infrared targets clearly but also provides richer texture information, significantly improving the overall visual quality. Therefore, infrared and visible image fusion technology is widely applied in many fields, like tracking [9], surveillance [10–12], and target detection [13,14].

Existing multi-source image fusion methods can be broadly categorized into traditional image fusion methods and deep learning-based image fusion methods. Traditional fusion methods primarily include multi-scale transform-based fusion methods [15,16], sparse representation-based fusion methods [17,18], subspace-based fusion frameworks [19,20], and saliency-based fusion methods [21,22]. These methods perform well in specific scenarios but often require adjustments to fusion strategies in response to environmental factors such as illumination and background variations. Consequently, traditional image fusion methods exhibit limited generalizability and require considerable time for the manual design of appropriate fusion workflows [23].

In recent years, deep learning-based methods for infrared and visible image fusion have attracted extensive attention within the image fusion community. Compared with traditional image fusion methods, deep learning-based approaches offer greater adaptability and enable end-to-end optimization, thereby reducing their reliance on manually designed fusion rules [3]. By designing effective network architectures and loss functions, these methods guide models to generate fused images enriched with more comprehensive visual information. Moreover, the robustness of these algorithms can be further enhanced by expanding the quantity and diversity of training datasets. Current deep learning-based image fusion algorithms primarily focus on addressing three critical challenges in the fusion process: feature extraction, feature fusion, and image reconstruction. Based on the adopted network architectures, deep learning-based infrared and visible image fusion methods can be categorized into three main categories: convolutional neural network (CNN)-based fusion methods [24–27], generative adversarial network (GAN)-based fusion methods [28–32] and autoencoder (AE)-based fusion methods [33–45].

CNN-based image fusion methods leverage convolutional kernels for effective local and multi-level feature extraction. In [46], an unsupervised end-to-end network was proposed to estimate the importance of source images, enabling adaptive information retention. Tang et al. [27] introduced SeAFusion, a semantic-aware real-time fusion network that connected image fusion with high-level vision tasks. However, CNN-based methods struggle with the lack of ground truth, as their performance heavily depends on loss functions and network design. To address this, GAN-based methods, such as [29], used adversarial generation to reduce the impact of loss functions and network architecture. Ma et al. [28] introduced GANMcC with multi-class constraints, which improved contrast and texture detail in fused images.

Despite providing a viable solution for image fusion, GAN-based methods face challenges such as high training complexity, instability during training and resource inefficiency. In contrast, AE-based image fusion methods circumvent these issues and offer advantages such as flexible fusion strategies, strong adaptability and robust performance, making them a more practical choice in certain applications. AE-based methods typically consist of an encoder and a decoder. Initially, a pre-trained encoder is employed to extract features from multi-source images. These extracted deep features are then fused using manually designed fusion strategies. Finally, a decoder is utilized to reconstruct the image from the fused features.

Currently, most AE methods primarily use CNNs for feature extraction, which inherently limits the network's ability to capture global features. As a result, some researchers have introduced Transformers into the design of feature extraction networks. The Transformer, being an effective network module for extracting long-distance dependencies, performs well in capturing global information in images [47]. However, since the Transformer was originally designed for natural language processing, this isolated usage increases global information but can lead to the loss of local information [48]. Consequently, the integration of CNNs and Transformers has emerged as a promising research direction, aiming to combine the advantages of both local and global feature representations. However, how to effectively coordinate the interaction between these heterogeneous features remains a critical challenge. Some studies, such as CDDFuse [41,49] employ a dual-branch architecture combining a CNN and Transformer to capture both local and global features independently. Other approaches apply a Transformer as a fusion module to establish interactions between different modalities or feature sources [50,51]. In addition, some recent studies [52,53] further exploit multi-scale receptive field features by feeding them into different Transformer modules, aiming to enrich the representation diversity. Despite these advances, most existing methods treat features of different spatial scales or receptive fields independently, without explicitly modeling the global structural relationships across feature scales and hierarchies. This may result in ineffective fusion of texture and structure information, compromising the visual quality and structural consistency of the fused image.

To address these challenges, this study proposes a progressive network architecture that first employs CNNs to extract hierarchical and scale-diverse local features and subsequently leverages a Transformer to model global dependencies across the resulting representations. This gradual integration allows the Transformer module to further process the extracted features, capturing both fine-grained details and long-range contextual relationships. Specifically, the module performs spatial-domain and scale-aware transformations to integrate complementary information while preserving spatial structural cues. By exploiting the self-attention mechanism, it facilitates the modeling of structural dependencies among important regions at different scales, thereby enhancing global contextual representation. Not only can it extract global information from features with different receptive fields, but it also lays the foundation for the alignment of multi-scale features in the subsequent fusion process. Additionally, a significance-preserving fusion module is introduced. This module dynamically adjusts the fusion weights at the feature level using a spatial selection mechanism to generate significance attention maps, which preserve the effective information in the salient regions of the source images. Furthermore, we propose a significance similarity metric. It adaptively computes the global structural correlation of images based on the extracted global features and calculates the similarity loss of salient regions using the attention maps from the fusion module. This approach ensures the retention of effective information in important regions, thereby meeting the requirements of multimodal image fusion tasks in complex scenes.

The central distinction of SiPFusion lies in the progressive coupling of these stages. The network first uses convolution-based multi-scale local feature extraction to organize fine textures, object boundaries, and intermediate-scale structures and then applies Transformer-based global feature extraction to model long-range contextual dependencies on the enriched local representations. This progressive local-to-global pathway enables the global branch to benefit from reliable local details rather than operating on isolated or single-scale features. The significance-preserving fusion module further generates spatial selection masks for the visible/infrared local and global feature branches, allowing adaptive feature selection across both modalities and representation levels. Moreover, the local selection masks and global features are reused in the significance similarity loss so that the feature-level fusion decisions are explicitly linked with output-level salient-region preservation and global structural consistency.

The main contributions of this paper can be concluded as follows:

1. We propose a fusion framework based on progressive feature extraction and significance preserving for the task of visible and infrared image fusion. Unlike existing fusion strategies, this framework explicitly considers the adaptive integration of multi-scale local information with the globally structured representations of multi-scale local features, rather than merely reconstructing fused data based on the global correlations of local features. This strategy enables comprehensive cross-scale feature interaction and improves global contextual representation, facilitating the retention of effective information in important regions.
2. A significance-preserving fusion module is employed to adaptively fuse the extracted features. In this module, a spatial selection mechanism is used to perform adaptive weighting between features from different receptive fields of the multimodal inputs, improving the network's capability to merge the most critical spatial context areas efficiently.
3. We designed a significance similarity metric in the loss function, which guides the global and salient region similarity of the fused image based on intermediate features. This improves structural consistency and enhances the perceptual quality of the fused image.

The rest of this article is organized as follows. In Section 2, we will review related works, including AE-based image fusion methods and Transformer-based image fusion methods. In Section 3, we will present our proposed method in detail. Section 4 will present experimental results, including experimental settings and both qualitative and quantitative comparisons, along with ablation studies. Section 5 will provide a conclusion and summary of the paper.

2. Related Works

2.1. AE-Based Image Fusion Methods

Autoencoder (AE)-based fusion methods have attracted considerable attention due to their powerful feature representation capabilities and end-to-end training flexibility. In the image fusion process, the autoencoder extracts and compresses the features of both the infrared and visible images, capturing their key characteristics. These features are then fused to generate a combined feature representation. Specifically, infrared images usually provide thermal radiation information in low-light or nighttime environments, while visible images contain rich visual details under normal lighting conditions.

In the encoding networks of these methods, various architectural enhancements have been proposed to improve feature extraction. For instance, DenseFuse [33], NestFuse [34], and RFN-Nest [35] integrated dense connections, nested skip paths, and residual dense blocks, respectively, to enhance feature reuse and gradient flow. More recent designs, such

as Res2Fusion [36] and MTD Fusion [37], employed Res2Net modules and triple-dense structures to further strengthen multi-scale feature representation.

To better capture image information for varying frequency and modality distributions, decomposition-based strategies were introduced [54]. DIDFuse [38] separated source images into high-frequency and low-frequency components, while DRF [39] decomposed inputs into scene-relevant and modality-specific components. MACCNet [40] extended this idea with a dual-branch encoder combining multi-scale attention and cross-convolution to extract complementary features from infrared and visible modalities.

In parallel, recent works have explored two-stage training strategies to improve fusion performance by decoupling different learning objectives. For instance, FTSFN [55] employs a two-stage fusion strategy in which modality-specific features are first transferred and then enhanced through a supplementary module. This separation allows the network to better preserve complementary details from both infrared and visible images. Similarly, DANet [56] adopts a two-stage training paradigm that integrates a diffusion model into an autoencoder framework. Although these methods demonstrate impressive results, the staged training schemes often increase the training complexity and may suffer from suboptimal global-local interaction due to their sequential optimization process. In contrast, our method adopts a progressive network that unifies local and global feature modeling, enabling more coordinated feature modeling across modalities and scales.

2.2. Transformer-Based Image Fusion Methods

A Transformer is a deep learning architecture based on the self-attention mechanism that completely discards recurrent and convolutional structures. By adopting self-attention, the Transformer enables parallel computation, significantly improving training efficiency. Due to their highly parallelizable computation, strong ability to model long-range dependencies, and flexible input–output structure, Transformer-based image fusion methods have attracted increasing attention.

To overcome the limited receptive field of convolutional networks, several studies have incorporated Transformers into image fusion frameworks. THFuse [57] integrates Transformer blocks into a CNN-based pipeline to enhance the modeling of global and local structures. STFNet [42] replaces handcrafted fusion strategies with a Transformer-based feature fusion module that explicitly models contextual relationships. DATFuse [58] employs a dual-attention Transformer to preserve long-range complementary information across modalities, whereas YDTR [59] adopts a Y-shaped dual-branch architecture with a dynamic Transformer to extract thermal and texture information. Trans2Fuse [60] further combines self-supervised learning with a Transformer-based fusion network to improve feature representation.

Recent studies have explored more specialized Transformer-based fusion designs. ATFuse [61] employs alternating cross-attention to capture inter-modal discrepancies and complementary information. The one-stream one-stage pure Transformer framework [62] integrates feature extraction and fusion into a unified Transformer pipeline. D2Fusion [63] explicitly decomposes and fuses high- and low-frequency components in both the spatial and frequency domains. In addition, recent Mamba-based approaches [64] have been investigated as alternative mechanisms for modeling long-range dependencies.

Existing hybrid fusion frameworks can be further distinguished by how they organize local and global representations and establish cross-modal interactions. CDDFuse [41] employs parallel Lite-Transformer and invertible CNN branches to represent low-frequency shared information and high-frequency modality-specific information with a correlation-driven objective. SwinFusion [65] models cross-domain interactions using Swin Transformer blocks. CrossFuse [66] introduces cross-attention between deep modality features

extracted by an autoencoder, whereas PPIFuse [67] combines modality-specific physical priors with multi-scale cross-attention. TransFusion [68] integrates cross-modal channel perception and deep-feature-guided spatial attention with a feature mutual-information objective. From the perspective of degradation-robust representation learning, URFusion [69] separates intrinsic content modeling from appearance representation assignment.

Taken together, these methods advance infrared and visible image fusion through representation decomposition, cross-modal interaction, physical-prior modeling, and degradation-robust learning. Building on these advances, SiPFusion approaches fusion from the perspective of cross-stage coordination. Its progressive local-to-global feature modeling provides representations for significance-aware feature selection, while the resulting intermediate masks and global features are further employed by S-SIM to guide output-level information preservation. This mechanism establishes a direct connection among feature extraction, feature fusion, and learning supervision.

3. The Proposed Method

In this part, the proposed significance-preserving progressive network named SiPFusion will be introduced in detail, including the architecture of the network, the encoder, the significance-preserving fusion module, the decoder and the loss function.

3.1. Network Framework

The overall architecture of the proposed SiPFusion network is shown in Figure 1.

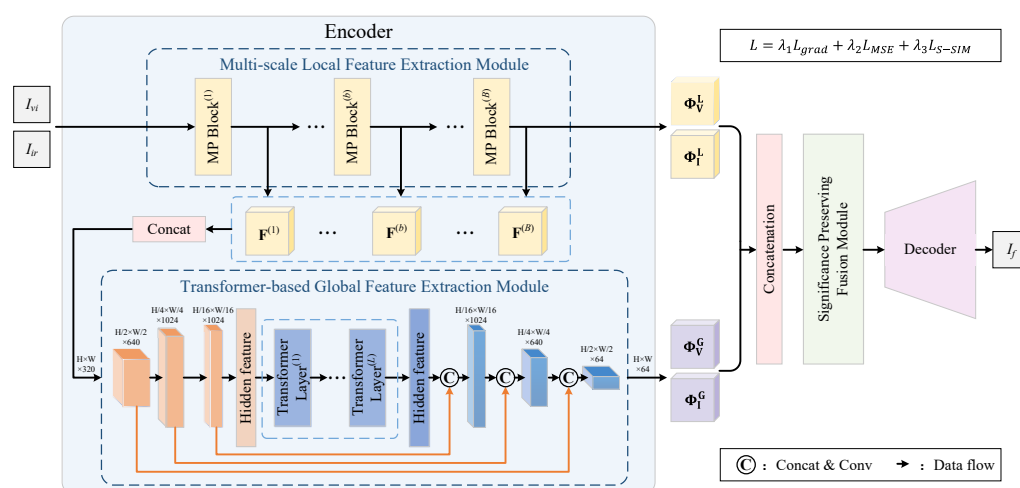


Figure 1. The overall framework of the proposed SiPFusion. The model contains a multi-scale local feature extraction (MLFE) module and a Transformer-based global feature extraction (TGFE) module. The multi-scale local feature extraction module utilizes MP blocks to extract local features from the source images. The Transformer-based global feature extraction module consists of several Transformer layers, which capture global dependencies based on the local features. The features extracted through the progressive structure are then fused using the fusion module, and finally, the fused image is reconstructed through the decoder.

The network consists of three main components: the encoder, the fusion module, and the decoder. Given an RGB visible image $I_{vi}^{RGB} \in \mathbb{R}^{H \times W \times 3}$, we first convert it to the YCbCr color space and use only its luminance channel as the visible input to the fusion network. For simplicity, this single-channel luminance component is denoted as $I_{vi} \in \mathbb{R}^{H \times W \times 1}$, while the infrared input is denoted as $I_{ir} \in \mathbb{R}^{H \times W \times 1}$. Both I_{ir} and I_{vi} are passed through the multi-scale local feature extraction module to obtain feature Φ_I^L, Φ_V^L , which can be expressed as

$$\Phi_I^L = \text{MLFE}(I_{ir}), \quad (1)$$

$$\Phi_V^L = \text{MLFE}(I_{vi}), \quad (2)$$

where Φ_I^L and Φ_V^L represent the final local feature representations of the infrared and visible branches used for subsequent feature fusion. During the MLFE process, the outputs of the B multi-scale perception (MP) blocks are concatenated to form modality-specific intermediate multi-scale feature maps, denoted as $\mathbf{F}_{ir}$ and $\mathbf{F}_{vi}$ for the infrared and visible branches, respectively. These concatenated feature maps are then fed into the Transformer-based global feature extraction module for further global dependency modeling, which can be expressed as

$$\Phi_I^G = \text{TGFE}(\mathbf{F}_{ir}), \quad (3)$$

$$\Phi_V^G = \text{TGFE}(\mathbf{F}_{vi}), \quad (4)$$

where TGFE indicates the global feature extraction operation of this module, which consists of L Transformer layers. This operation ultimately produces the feature tensor that contains both local and global information. Then, the four feature maps generated by the two modules are concatenated, which can be denoted as

$$\Phi = \text{Concat}(\Phi_I^L, \Phi_V^L, \Phi_I^G, \Phi_V^G), \quad (5)$$

where Concat refers to the concatenation operation. Afterward, the concatenated feature map Φ is processed through the significance-preserving fusion (SPF) module to obtain the final fused feature $\hat{\mathbf{F}}$. This procedure can be formulated as

$$\hat{\mathbf{F}} = \text{SPF}(\Phi) \quad (6)$$

Finally, the fused feature is reconstructed to the original size using a decoder, resulting in the final fused image I_f :

$$I_f = \psi_D(\hat{\mathbf{F}}), \quad (7)$$

where ψ_D represents the decoder block.

3.2. Progressive Structure-Based Encoder Network

Considering that both local salient target information and the global correlations in infrared and visible images play a crucial role in understanding the image scene and maintaining global structural consistency, we designed a network architecture in the encoder that combines a multi-scale local feature extraction module with a Transformer-based global feature extraction module. This approach further enhances the network's ability to simultaneously integrate both the detailed features and overall characteristics of the source images.

3.2.1. Multi-Scale Local Feature Extraction Module

As the local feature extractor of SiPFusion, the MLFE module aims to capture local contextual information at multiple receptive-field scales, with particular emphasis on salient targets and fine texture details. Given a source image $I \in \{I_{ir}, I_{vi}\}$, a shallow 3×3 convolution first projects the input into a feature representation with a base channel width of $C = 64$. The resulting feature is then progressively processed by B sequential MP blocks. Instead of using only the output of the last block, MLFE retains the output of each MP block and concatenates them along the channel dimension to aggregate features extracted at different depths. This process is formulated as

$$\begin{aligned}
\mathbf{F}^{(0)} &= \mathcal{C}_{3 \times 3}(I), \\
\mathbf{F}^{(i)} &= \text{MP}\left(\mathbf{F}^{(i-1)}\right), \quad i = 1, 2, \dots, B, \\
\mathbf{F} &= \text{Concat}\left(\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \dots, \mathbf{F}^{(B)}\right),
\end{aligned} \tag{8}$$

where $\mathcal{C}_{3 \times 3}(\cdot)$ denotes the shallow feature projection, $\mathbf{F}^{(i)}$ represents the output of the i -th MP block, and $\mathbf{F}$ denotes the concatenated multi-scale local feature for a single modality branch. When the input image is I_{ir} or I_{vi} , this concatenated feature is denoted as $\mathbf{F}_{ir}$ or $\mathbf{F}_{vi}$, respectively, and is used as the input to TGFE.

As illustrated in Figure 2, each MP block contains three parallel convolutional branches with kernel sizes of 1×1 , 3×3 , and 5×5 , respectively. Their output channels are set to C , $C/2$, and $C/4$. Since $C = 64$ in our experiments, the three branches produce 64, 32, and 16 channels, respectively. All branches use a stride of 1 with appropriate padding to preserve the spatial resolution of the feature maps. The outputs of the three branches are concatenated along the channel dimension, producing a 112-channel intermediate feature. A subsequent 1×1 convolution projects this feature back to 64 channels, after which a residual connection is applied between the projected feature and the block input. Therefore, each MP block preserves the spatial and channel dimensions of its input and produces $\mathbf{F}^{(i)} \in \mathbb{R}^{H \times W \times 64}$. In our implementation, B is set to 5. Concatenating the outputs of the five MP blocks therefore produces $\mathbf{F} \in \mathbb{R}^{H \times W \times (B \cdot C)} = \mathbb{R}^{H \times W \times (5 \cdot 64)} = \mathbb{R}^{H \times W \times 320}$. The resulting $H \times W \times 320$ feature tensor is subsequently fed into the global feature extraction stage.

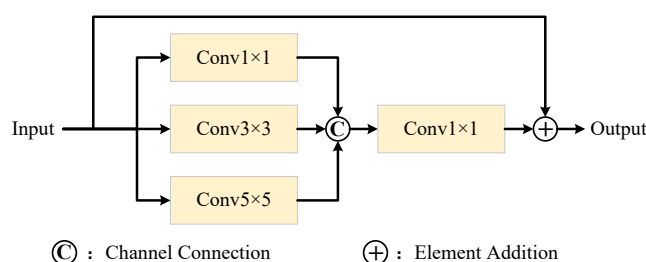


Figure 2. The architecture of MP block.

3.2.2. Transformer-Based Global Feature Extraction Module

The MLFE module is a CNN-based local extractor and thus has inherent locality, which limits its ability to model global dependencies across the entire image, especially for distant targets. To enhance global dependency learning, we further introduce a TGFE module.

In our implementation, the concatenated MLFE features $\mathbf{F}_{ir}$ and $\mathbf{F}_{vi}$ are fed into TGFE to extract global representations for the infrared and visible branches, respectively. TGFE starts with a downsampling pyramid that progressively reduces the spatial resolution while expanding the receptive field, producing three feature maps at resolutions $H/2$, $H/4$, and $H/16$ (with channel sizes 640, 1024, and 1024, respectively). The $H/8$ level is intentionally omitted to limit further growth in the number of spatial tokens used for global self-attention. During decoding, the $H/16$ feature is directly upsampled to $H/4$ and fused with the corresponding skip feature, while intermediate-scale spatial information is retained through the $H/4$ and $H/2$ skip connections. The coarsest feature map is then projected by patch embedding to a hidden dimension D (set to 128 in our configuration) and reshaped into a token sequence of length $N_{tok} = (H/16) \cdot (W/16)$. The TGFE module contains L Transformer layers (LayerNorm, multi-head attention, and MLP) to capture long-range dependencies for global contextual modeling.

After Transformer encoding, the token sequence is reshaped back to a 2D feature map of size $(H/16, W/16, D)$ and decoded by progressive upsampling with three-stage skip fusion. Specifically, at each stage, the upsampled feature map is concatenated with the corresponding downsampling-stage feature map at the same spatial resolution and then projected by convolution (denoted as “C: Concat & Conv” in Figure 1). The decoding path restores the resolution in the order $(H/16, W/16) \rightarrow (H/4, W/4) \rightarrow (H/2, W/2) \rightarrow (H, W)$ and finally outputs a full-resolution global feature map $\Phi^G \in \mathbb{R}^{H \times W \times 64}$ for each modality (Φ_I^G and Φ_V^G).

3.3. Significance-Preserving Fusion Module

In terms of fusion strategy, we adopt a significance-preserving mechanism to maintain the salient regions in the feature maps of the multi-source images, thereby improving the network’s capability to concentrate on the most informative spatial context regions. The structure of SPF module is illustrated in Figure 3.

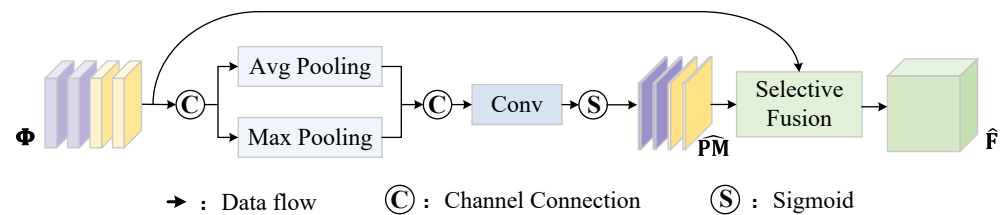


Figure 3. The architecture of SPF Block.

Initially, as shown in Equation (5), we concatenate the four feature maps of the infrared and visible images obtained from the progressive structure. Then, operations are performed on Φ through average pooling and max pooling at the feature channel level:

$$\begin{aligned} \mathbf{PM}_{avg} &= \mathcal{P}_{avg}(\Phi), \\ \mathbf{PM}_{max} &= \mathcal{P}_{max}(\Phi), \end{aligned} \quad (9)$$

where $\mathcal{P}_{avg}(\cdot)$ and $\mathcal{P}_{max}(\cdot)$ represent the average pooling operation and the max pooling operation, respectively, while $\mathbf{PM}_{avg}$ and $\mathbf{PM}_{max}$ are their corresponding saliency feature descriptors. To facilitate the interaction of salient information between different source features, the saliency feature descriptors $\mathbf{PM}_{avg}$ and $\mathbf{PM}_{max}$ are aggregated into a single significance feature descriptor. This aggregated descriptor is then transformed into the corresponding N_{map} significance attention maps through a convolution operation $\mathcal{C}^{2 \rightarrow N_{map}}(\cdot)$, which can be described as

$$\widetilde{\mathbf{PM}} = \mathcal{C}^{2 \rightarrow N_{map}}(\text{Concat}(\mathbf{PM}_{avg}, \mathbf{PM}_{max})), \quad (10)$$

where $N_{map} = 4$, and $\widetilde{\mathbf{PM}}$ represents the four significance attention maps generated for the local and global features of the two modalities. Specifically, the four attention maps are denoted as $\widetilde{\mathbf{PM}}_V^L$, $\widetilde{\mathbf{PM}}_V^G$, $\widetilde{\mathbf{PM}}_I^L$, and $\widetilde{\mathbf{PM}}_I^G$, which correspond one-to-one to Φ_V^L , Φ_V^G , Φ_I^L , and Φ_I^G , respectively. The Sigmoid activation function is then applied to each attention map to obtain its corresponding significance selection mask:

$$\widehat{\mathbf{PM}}_m^s = \rho(\widetilde{\mathbf{PM}}_m^s), \quad m \in \{V, I\}, \quad s \in \{L, G\}, \quad (11)$$

where m denotes the visible or infrared modality, s denotes the local or global feature level, and $\rho(\cdot)$ denotes the Sigmoid activation function.

The four extracted features are weighted by their corresponding selection masks and then summed to obtain the fused feature tensor:

$$\hat{\mathbf{F}} = \widehat{\mathbf{PM}}_V^L \odot \Phi_V^L + \widehat{\mathbf{PM}}_V^G \odot \Phi_V^G + \widehat{\mathbf{PM}}_I^L \odot \Phi_I^L + \widehat{\mathbf{PM}}_I^G \odot \Phi_I^G, \quad (12)$$

where $\odot$ denotes element-wise multiplication. Each selection mask weights only its corresponding feature map, and all four masks participate in the adaptive feature fusion performed by SPF.

Although SPF adopts an attention-based weighting form, its purpose is not generic feature refinement but adaptive multi-branch feature fusion. Specifically, SPF receives the visible local feature, visible global feature, infrared local feature, and infrared global feature and generates branch-specific spatial selection maps to adaptively select complementary information across both modality and representation level. Moreover, the local selection maps corresponding to the visible and infrared local features are reused by the SRS term in S-SIM, thereby linking feature-level fusion with output-level salient-region preservation. This design makes SPF a fusion-oriented selection module rather than a standalone attention-refinement block.

3.4. Decoder

The decoder performs a decoding operation on the features fused by the SPF module. The decoder consists of 1×1 and 3×3 convolution layers, a Leaky ReLU activation function, followed by another 3×3 convolution layer with a Sigmoid activation function. This process ultimately restores the features to their original spatial dimensions and outputs a single-channel fused luminance image I_f . During inference, I_f is combined with the Cb and Cr channels retained from the original visible image. The resulting YCbCr image is then converted back to the RGB color space to obtain the final color fused image.

3.5. Loss Function

The loss function plays a guiding role during the training phase of the network. In this paper, we propose a hybrid loss function, which includes gradient loss, mean squared error (MSE) loss and a significance similarity (S-SIM) loss that we propose. The formula for the total loss function is as follows:

$$L = \lambda_1 L_{grad} + \lambda_2 L_{MSE} + \lambda_3 L_{S-SIM}, \quad (13)$$

where λ_1, λ_2 , and λ_3 are weighting coefficients used to balance the numerical contributions of the gradient, MSE, and S-SIM losses, respectively.

Gradient loss is calculated by computing the maximum gradient between the fused image and the two source images. The gradient values at corresponding locations are subtracted, and the squared differences are averaged across all pixels. Gradient loss helps the fused image retain the detail and texture information from the source images. The formula can be expressed as

$$L_{grad} = \frac{1}{HW} \sum_{i=1}^{HW} \left(\nabla I_f(i) - \max(\nabla I_{ir}(i), \nabla I_{vi}(i)) \right)^2, \quad (14)$$

where $\nabla I(i)$ represents the gradient of the image at pixel i , and HW denotes the total number of pixels in the image.

MSE loss can evaluate the fusion performance by calculating the squared differences between the fusion image and the source images at each pixel position, which has certain advantages for network training. It can be expressed as

$$L_{MSE} = \alpha_1 \left(\frac{1}{HW} \sum_{i=1}^{HW} (I_f(i) - I_{ir}(i))^2 \right) + \alpha_2 \left(\frac{1}{HW} \sum_{i=1}^{HW} (I_f(i) - I_{vi}(i))^2 \right), \quad (15)$$

where α_1, α_2 represent the balancing parameters for the MSE of the infrared and visible images. In order to ensure that the fused image retains as much information as possible from both source images, we assume that the contributions of the two source images are equal. Therefore, both α_1 and α_2 are set to 1.

In order to better preserve the global structural similarity and the effective information of the salient regions in the image, we designed the S-SIM loss. The S-SIM loss function primarily consists of two components: the salient region similarity (SRS) loss and the global structural similarity (GSS) loss. Its formulation is as follows:

$$L_{S-SIM} = \beta_1 L_{SRS} + \beta_2 L_{GSS}, \quad (16)$$

where β_1 and β_2 are weighting coefficients used to balance the numerical contributions of the SRS and GSS terms, respectively.

The goal of the SRS loss is to preserve informative intensity content in salient local regions. Among the four selection masks generated by SPF, $\widehat{\mathbf{PM}}_V^L$ and $\widehat{\mathbf{PM}}_I^L$, which correspond to the visible and infrared local features Φ_V^L and Φ_I^L , respectively, are directly reused to construct the spatial source-selection weights $\mathbf{W}_1$ and $\mathbf{W}_2$. The weights are defined as

$$\begin{cases} \mathbf{W}_1 = \delta(\widehat{\mathbf{PM}}_V^L - \widehat{\mathbf{PM}}_I^L), \\ \mathbf{W}_2 = 1 - \mathbf{W}_1, \end{cases} \quad (17)$$

where $\delta(\cdot)$ denotes the step function. During training, the step function is implemented using the differentiable approximation $\delta(x) \approx \sigma((x - \text{thr})/\tau)$, where $\sigma(z) = 1/(1 + \exp(-z))$ is the standard sigmoid function, thr is the decision threshold, and τ controls the sharpness of the approximation. We set $\text{thr} = 0$ and $\tau = 0.1$ in all experiments.

Using these spatial weights, the SRS loss is calculated as

$$L_{SRS} = \frac{1}{HW} \sum_{i=1}^{HW} [\mathbf{W}_1 |I_f(i) - I_{vi}(i)| + \mathbf{W}_2 |I_f(i) - I_{ir}(i)|], \quad (18)$$

where HW denotes the total number of pixels. Thus, the fused image is adaptively constrained toward the source modality assigned a larger local selection weight at each spatial location.

To better preserve effective scene-related information such as global positions, we further design the GSS loss as a global scale similarity metric loss function. Based on the Φ_V^G and Φ_I^G obtained from the Transformer-based feature extraction module, we calculate their means μ_{vi}, μ_{ir} to obtain the weights $\mathbf{W}_3$ and $\mathbf{W}_4$ for the global structural similarity. The specific calculation formula is as follows:

$$\begin{cases} \mu_{vi} = \text{MEAN}(\Phi_V^G) \\ \mu_{ir} = \text{MEAN}(\Phi_I^G) \end{cases}, \quad (19)$$

$$\begin{cases} \mathbf{W}_3 = \frac{\mu_{vi}}{\mu_{vi} + \mu_{ir}} \\ \mathbf{W}_4 = \frac{\mu_{ir}}{\mu_{vi} + \mu_{ir}} \end{cases}, \quad (20)$$

where $\text{MEAN}(\cdot)$ denotes the global averaging operation. The normalized weights $\mathbf{W}_3$ and $\mathbf{W}_4$ are used to adaptively balance the visible- and infrared-specific structural similarity terms according to the global responses of Φ_V^G and Φ_I^G . The structural similarity is calculated as follows:

$$\omega(X, F) = \frac{(2\mu_X\mu_F + C_1)(2\sigma_{XF} + C_2)}{(\mu_X^2 + \mu_F^2 + C_1)(\sigma_X^2 + \sigma_F^2 + C_2)}, \quad (21)$$

where $\omega(X, F)$ represents the calculation of the structural similarity between source image X and fused image F , μ_X refers to the mean of the image X , and σ_X and σ_{XF} represent the variance and covariance of the images, respectively.

Since a higher SSIM value indicates better consistency between the fused image and the source image, the loss is computed as $1 - \text{SSIM}$ in the loss function. Finally, the GSS loss is calculated by weighting the loss based on the global features obtained from the weight map.

$$L_{\text{GSS}} = \mathbf{W}_3 \left[1 - \omega(I_{vi}, I_f) \right] + \mathbf{W}_4 \left[1 - \omega(I_{ir}, I_f) \right]. \quad (22)$$

4. Experiments

In this section, we first introduce the experimental settings of this paper, including the datasets, evaluation metrics, and details of the experimental process. Next, we present the comparison algorithms and the results of both qualitative and quantitative analyses. Finally, we elaborate on the ablation experiments.

4.1. Experimental Settings

4.1.1. Dataset Description

The infrared and visible image fusion experiments conducted in this paper used three publicly available datasets: MSRS [26], RoadScene [24], and TNO [70]. The MSRS and RoadScene datasets contain pairs of infrared and visible images from various daytime and nighttime road scenes, while the TNO dataset includes pairs from a broader range of scenes, such as military environments.

Specifically, 1294 image pairs from the MSRS dataset were used for training, while the remaining 150 pairs were retained for evaluation. The trained model was further evaluated on 25 image pairs from RoadScene and 25 image pairs from TNO. No fine-tuning was performed during testing.

4.1.2. Evaluation Metrics

To evaluate the fusion results from complementary perspectives, we employed five metrics [71]: mean squared error (MSE), peak signal-to-noise ratio ($PSNR$), fusion artifacts ($N^{AB/F}$), correlation coefficient (CC), and multi-scale structural similarity ($MS\text{-}SSIM$). In Section 4.1.2, MSE refers to the evaluation metric rather than the training loss term L_{MSE} defined in Section 3.5. The evaluation MSE was computed as a dual-reference measure between the fused image and the two source images. All metrics were computed on single-channel images. The symbols I_{vi} , I_{ir} , and I_f denote the visible luminance image, infrared image, and fused luminance image, respectively. Each metric was first calculated for every test image pair and then averaged over the corresponding test set.

For the metrics requiring normalized intensities, the images were mapped from $[0, 255]$ to $[0, 1]$ as

$$\bar{I}_x = \frac{I_x}{255}, \quad x \in \{vi, ir, f\}. \quad (23)$$

The *MSE* measures the average pixel-wise discrepancy between the fused image and the two source images. Since infrared-visible image fusion does not provide a ground-truth fused image, the pairwise errors between the fused image and each source image were first calculated as

$$MSE_{x,f} = \frac{1}{HW} \sum_{i=1}^{HW} \left(\bar{I}_f(i) - \bar{I}_x(i) \right)^2, \quad x \in \{vi, ir\}, \quad (24)$$

$$MSE = \frac{MSE_{vi,f} + MSE_{ir,f}}{2}, \quad (25)$$

where H and W denote the image height and width, respectively. A lower *MSE* indicates a smaller average pixel-wise discrepancy between the fused image and the two source images.

The *PSNR* was calculated from the dual-reference *MSE* defined above:

$$PSNR = 10 \log_{10} \left(\frac{I_{\max}^2}{MSE} \right), \quad (26)$$

where $I_{\max} = 1$ denotes the maximum possible intensity value after normalization to $[0, 1]$. A higher *PSNR* indicates a smaller average pixel-level distortion between the fused image and the two source images.

The artifact measure $N^{AB/F}$ evaluates the amount of artificial edge information introduced into the fused image. It is defined as

$$N^{AB/F} = \frac{\sum_{i=1}^H \sum_{j=1}^W AM(i,j) \left[\left(1 - Q^{vi,f}(i,j) \right) \omega^{vi}(i,j) + \left(1 - Q^{ir,f}(i,j) \right) \omega^{ir}(i,j) \right]}{\sum_{i=1}^H \sum_{j=1}^W \left[\omega^{vi}(i,j) + \omega^{ir}(i,j) \right]}, \quad (27)$$

where $Q^{vi,f}$ and $Q^{ir,f}$ denote the edge-information preservation measures between the fused image and the visible and infrared source images, respectively. The artifact indicator is defined as

$$AM(i,j) = \begin{cases} 1, & g_f(i,j) > g_{vi}(i,j) \text{ and } g_f(i,j) > g_{ir}(i,j), \\ 0, & \text{otherwise,} \end{cases} \quad (28)$$

where g_{vi} , g_{ir} , and g_f denote the edge strengths of the visible, infrared, and fused images, respectively, while ω^{vi} and ω^{ir} denote the corresponding edge-importance weights. A lower $N^{AB/F}$ indicates fewer artifacts introduced during fusion.

The *CC* evaluates the linear correlation between the fused image and the two source images:

$$CC_{x,f} = \frac{\sum_{i=1}^{HW} (I_x(i) - \mu_x) (I_f(i) - \mu_f)}{\sqrt{\sum_{i=1}^{HW} (I_x(i) - \mu_x)^2 \sum_{i=1}^{HW} (I_f(i) - \mu_f)^2}}, \quad x \in \{vi, ir\}, \quad (29)$$

$$CC = \frac{CC_{vi,f} + CC_{ir,f}}{2}, \quad (30)$$

where μ_x and μ_f denote the mean intensities of the source and fused images, respectively. A higher *CC* indicates a stronger average linear correlation between the fused image and the two source images.

The standard pairwise *MS-SSIM* measures structural similarity between two images across multiple spatial scales. For a source image I_x and the fused image I_f , it is calculated as

$$\text{MSSSIM}(\bar{I}_x, \bar{I}_f) = \left[l_f(\bar{I}_x, \bar{I}_f) \right]^{w_j} \prod_{j=1}^J \left[c_j(\bar{I}_x, \bar{I}_f) s_j(\bar{I}_x, \bar{I}_f) \right]^{w_j}, \quad (31)$$

where $l_f(\cdot)$ denotes the luminance comparison at the coarsest scale, while $c_j(\cdot)$ and $s_j(\cdot)$ denote the contrast and structural comparisons at scale j , respectively. The symbol w_j denotes the standard scale weight [72].

Because infrared-visible image fusion involves two source images, the *MS-SSIM* reported in this paper is defined as the sum of the two pairwise scores:

$$\text{MS-SSIM} = \text{MSSSIM}(\bar{I}_{vi}, \bar{I}_f) + \text{MSSSIM}(\bar{I}_{ir}, \bar{I}_f). \quad (32)$$

Each pairwise MSSSIM value lies within $[0, 1]$; therefore, the summed dual-reference *MS-SSIM* reported in this paper lies within $[0, 2]$. A higher value indicates stronger overall structural similarity between the fused image and the two source images.

4.1.3. Implementation Details

Our experiments were conducted using the PyTorch 2.4.0 framework on an NVIDIA A800 GPU with 80 GB of memory. In the preprocessing stage, the training samples were randomly divided into 128×128 patches. The training process was run for a total of 60 epochs. The batch size was set to 128. The optimizer used was Adam with the initial learning rate set to 10^{-4} . For the network hyperparameters, the number of MP blocks B was set to 5. The TGFE module consisted of $L = 8$ Transformer layers.

For the overall loss function in Equation (13), the weights were empirically selected through preliminary calibration experiments based on the observed numerical scales of the unweighted loss terms. Since the unweighted gradient loss was considerably smaller than the MSE and S-SIM losses, λ_1 was set to 50 to maintain an effective gradient-preservation constraint. The MSE and S-SIM terms were assigned weights of $\lambda_2 = 5$ and $\lambda_3 = 2$, respectively. For the S-SIM loss in Equation (16), the internal weights were set to $\beta_1 = 1$ and $\beta_2 = 5$ following the same scale-balancing principle. The resulting coefficient configuration was fixed for all reported experiments.

4.1.4. Comparison with State-of-the-Arts

We evaluated the proposed method on three datasets and compared it with 15 state-of-the-art methods, namely, one traditional method (MDLatLRR [16]), four CNN-based methods (FusionDN [24], ReCoNet [73], SeAFusion [27], PIAFusion [26]), one GAN-based method (TarDAL [31]), four AE-based methods (NestFuse [34], DIDFuse [38], DeFusion [74], CrossFuse [66]), two Transformer-based methods (SwinFusion [65], CDDFuse [41]), and three recent methods published in 2025 (PIIFuse [67], TransFusion [68], and URFusion [69]).

4.2. Qualitative Analysis

Figures 4–6 present qualitative comparisons between SiPFusion and 15 representative fusion methods on the MSRS, RoadScene, and TNO datasets, respectively. The highlighted regions were used to examine the preservation of scene intensity, salient targets, and local structural details.

As shown in Figure 4, several comparison methods exhibit uneven intensity distributions or local texture distortions in the green-boxed road region. CrossFuse presents noticeable brightness discontinuities, whereas the results produced by ReCoNet, DIDFuse,

and TransFusion appear relatively dark, making some low-light structures less distinguishable. In comparison, SiPFusion maintains a coherent road-region representation with smoother intensity transitions and fewer visually apparent high-frequency fluctuations. In the red-boxed low-light region, the tree branches remain recognizable, while the overall scene intensity is preserved consistently.

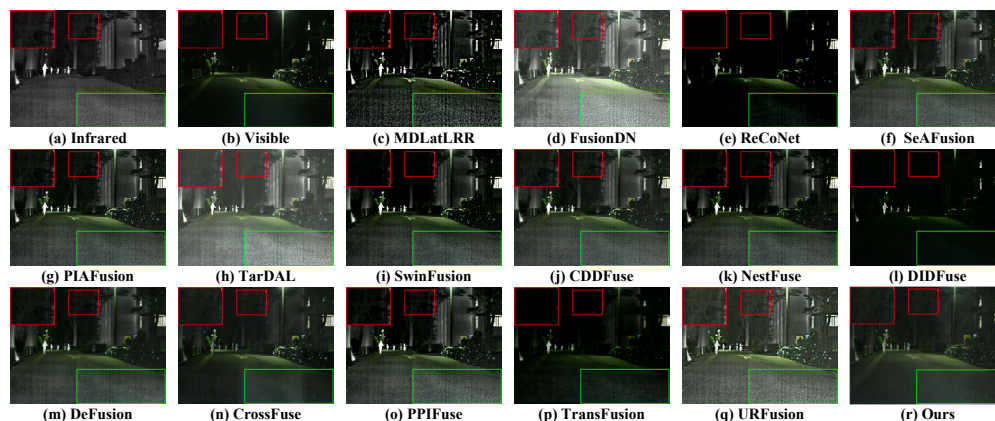


Figure 4. Qualitative comparison of SiPFusion with 15 state-of-the-art methods on MSRS dataset (01035N).

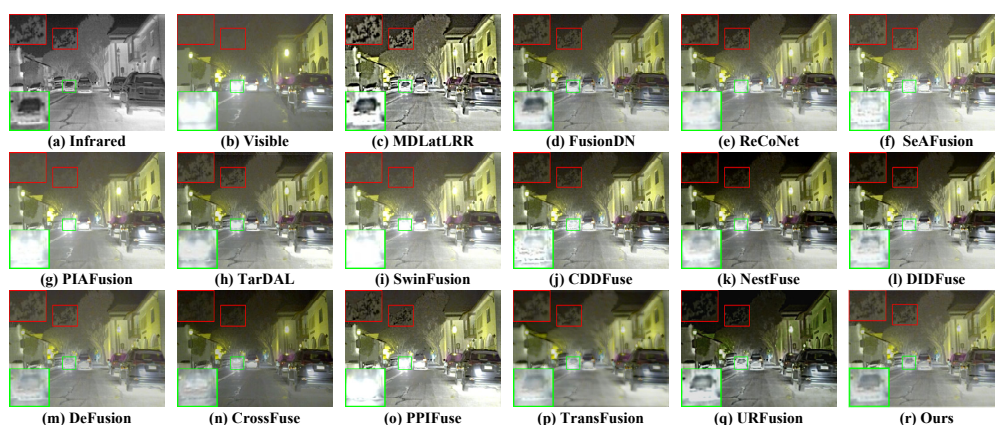


Figure 5. Qualitative comparison of SiPFusion with 15 state-of-the-art methods on RoadScene dataset (FLIR_09336).

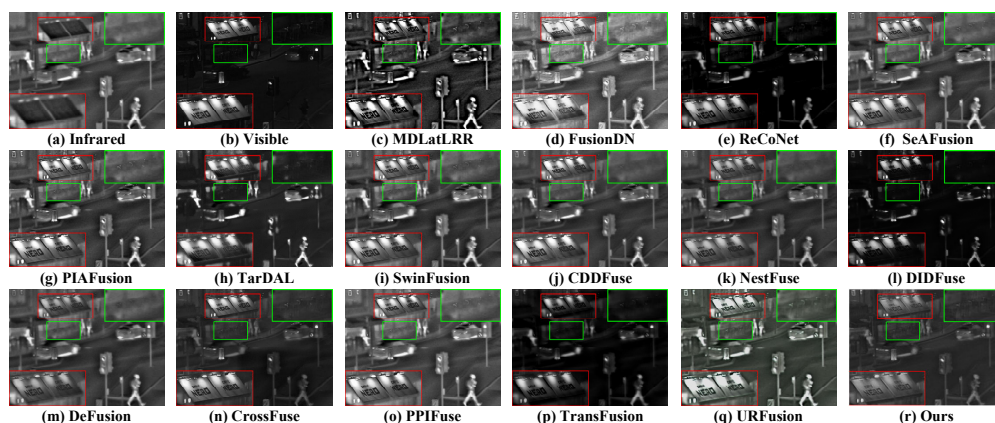


Figure 6. Qualitative comparison of SiPFusion with 15 state-of-the-art methods on TNO dataset (002).

In Figure 5, the green-boxed region contains a salient vehicle affected by surrounding headlight illumination. SiPFusion preserves the principal vehicle contours and local intensity information without introducing visually dominant halo-like artifacts. In the red-boxed region, which contains densely distributed tree branches, SiPFusion maintains

distinguishable branch structures without excessive sharpening. These observations indicate a balanced preservation of salient intensity information and local structural details.

In Figure 6, the green-boxed region highlights the railing structure in front of the store entrance. SiPFusion preserves the principal contours and individual segments of the railing, allowing these structures to remain distinguishable in the fused image. In the red-boxed region, recognizable patterns on the store sign are retained without visually conspicuous artifacts. The resulting image therefore maintains a balanced integration of visible texture and infrared intensity information.

To complement the visual observations, we evaluated the six highlighted regions in Figures 4–6. All calculations were performed on the luminance channel using identical region coordinates for all methods. Let $\mathcal{R}_k$ denote the k -th highlighted region, where $K = 6$. The local dual-source correlation coefficient and local dual-source structural similarity are defined as

$$\text{LCC} = \frac{1}{K} \sum_{k=1}^K \frac{CC(I_{vi}^{\mathcal{R}_k}, I_f^{\mathcal{R}_k}) + CC(I_{ir}^{\mathcal{R}_k}, I_f^{\mathcal{R}_k})}{2}, \quad (33)$$

$$\text{LSSIM} = \frac{1}{K} \sum_{k=1}^K \frac{SSIM(I_{vi}^{\mathcal{R}_k}, I_f^{\mathcal{R}_k}) + SSIM(I_{ir}^{\mathcal{R}_k}, I_f^{\mathcal{R}_k})}{2}. \quad (34)$$

As shown in Table 1, SiPFusion achieves the highest average LCC and competitive LSSIM performance. These results demonstrate that SiPFusion maintains strong aggregate consistency with the visible and infrared source images while competitively preserving local structural information.

Table 1. Localized quantitative comparison over the six highlighted regions in Figures 4–6. The best and second-best results are highlighted in bold and underlined, respectively. $\uparrow$ indicates that higher values are better.

Method	LCC $\uparrow$	LSSIM $\uparrow$
MDLatLRR	0.4921	0.4102
FusionDN	<u>0.5996</u>	0.5014
ReCoNet	<u>0.5537</u>	0.4054
SeAFusion	0.5595	0.6351
PIAFusion	0.5589	<u>0.6406</u>
TarDAL	0.5574	0.5515
SwinFusion	0.5560	0.5509
CDDFuse	0.5357	0.6240
NestFuse	0.5900	0.5308
DIDFuse	0.5854	0.4467
DeFusion	0.5563	0.6658
CrossFuse	0.5604	0.5825
PPIFuse	0.4859	0.5793
TransFusion	0.5669	0.5019
URFusion	0.5489	0.4896
SiPFusion	0.6249	0.6350

4.3. Quantitative Analysis

From Tables 2–4, SiPFusion achieves consistently strong quantitative performance against 15 competitive baselines, including recently published methods.

On MSRS, SiPFusion attains the best *MSE*, *PSNR*, *CC*, and *MS-SSIM* while remaining competitive on $N^{AB/F}$ compared with the lowest-artifact results.

On RoadScene, our method achieves the best *MSE*, *PSNR*, and *CC* and the second-best *MS-SSIM*, indicating improved fidelity and cross-modal consistency under challenging illumination and headlight glare conditions.

On TNO, SiPFusion again obtains the best *MSE*, *PSNR*, *CC*, and *MS-SSIM*, demonstrating robust generalization to diverse scenes. These improvements can be attributed to the complementary contributions of the proposed components.

First, the progressive local-to-global feature extraction (MLFE $\rightarrow$ TGFE) enhances cross-scale alignment and global dependency modeling, which reduces reconstruction error (lower *MSE*/higher *PSNR*) and benefits structure-related similarity (higher *MS-SSIM*).

Second, the SPF module performs spatial selection to emphasize informative regions and suppress irrelevant responses, leading to higher correlation with the sources (higher *CC*) and more stable structural preservation.

Third, the proposed S-SIM loss explicitly constrains salient-region fidelity and global structural consistency during training, further improving perceptual and structural quality.

Table 2. Quantitative comparisons on the MSRS dataset. In the table, **bold** is best, and underlined is second-best. $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively.

	<i>MSE</i> $\downarrow$	<i>PSNR</i> $\uparrow$	<i>N^{AB/F}</i> $\downarrow$	<i>CC</i> $\uparrow$	<i>MS-SSIM</i> $\uparrow$
MDLatLRR	0.04808	13.4962	0.1999	0.5782	1.0753
FusionDN	0.07565	11.2669	0.0726	0.6201	1.2853
ReCoNet	0.04242	14.0533	0.0315	0.5956	1.2580
SeAFusion	0.05386	13.2431	0.0427	0.6113	1.3421
PIAFusion	0.05871	12.9121	0.0800	0.6025	1.3292
TarDAL	0.05065	13.0972	0.0294	<u>0.6414</u>	<u>1.3599</u>
SwinFusion	0.05714	13.0542	0.0841	0.6017	1.3107
CDDFuse	0.05204	13.3920	0.0444	0.5973	1.3252
NestFuse	0.04865	13.8166	0.0505	0.6035	1.3291
DIDFuse	0.04618	13.6180	0.0284	0.6349	1.2646
DeFusion	<u>0.03679</u>	<u>14.9569</u>	0.0123	0.5836	1.3325
CrossFuse	0.04631	14.0026	0.0173	0.5494	1.3150
PPIFuse	0.05654	13.0708	0.0819	0.6016	1.3062
TransFusion	0.03681	14.6409	0.0073	0.6388	1.3402
URFusion	0.04952	13.1567	0.0681	0.5817	1.2904
Ours	0.02925	15.8090	<u>0.0100</u>	0.6877	1.4203

Table 3. Quantitative comparisons on the RoadScene dataset. In the table, **bold** is best, and underlined is second-best. $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively.

	<i>MSE</i> $\downarrow$	<i>PSNR</i> $\uparrow$	<i>N^{AB/F}</i> $\downarrow$	<i>CC</i> $\uparrow$	<i>MS-SSIM</i> $\uparrow$
MDLatLRR	0.05415	12.9204	0.1091	0.5106	1.0436
FusionDN	0.05530	13.0634	0.0256	<u>0.5642</u>	1.2831
ReCoNet	0.04870	13.6551	0.0292	0.5431	1.2973
SeAFusion	0.06635	12.1833	0.1086	0.5287	1.2555
PIAFusion	0.06719	12.2132	0.4545	0.5041	1.2788
TarDAL	0.04734	13.6449	0.0303	0.5064	1.2310
SwinFusion	0.08715	10.9973	0.1517	0.5312	1.2573
CDDFuse	0.05904	12.7755	0.0656	0.5424	1.2598
NestFuse	0.06694	12.0919	0.0264	0.5279	1.2455
DIDFuse	0.05021	13.2285	0.0468	0.5584	1.2597
DeFusion	<u>0.04131</u>	<u>14.4211</u>	0.0033	0.5258	1.2523
CrossFuse	0.08318	11.2965	0.0162	0.4628	1.1641
PPIFuse	0.07665	11.7237	0.0926	0.4921	1.1999
TransFusion	0.04176	14.1606	<u>0.0079</u>	0.5470	1.2637
URFusion	0.09629	10.3485	0.0175	0.5420	1.2572
Ours	0.03700	14.6731	0.0104	0.5826	<u>1.2952</u>

While SiPFusion does not obtain the lowest $N^{AB/F}$ values on the three datasets, it achieves strong fidelity-, correlation-, and structure-related performance while maintaining competitive artifact suppression, resulting in a favorable overall balance across the evaluated metrics.

Table 4. Quantitative comparisons on the TNO dataset. In the table, **bold** is best, and underlined is second-best. $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively.

	$MSE \downarrow$	$PSNR \uparrow$	$N^{AB/F} \downarrow$	$CC \uparrow$	$MS\text{-}SSIM \uparrow$
MDLatLRR	<u>0.05226</u>	13.1926	0.0887	0.4657	1.1515
FusionDN	0.05807	12.6321	0.0363	0.4834	1.2758
ReCoNet	0.05409	13.0883	0.0153	0.4780	1.3002
SeAFusion	0.07894	11.7012	0.0879	0.4695	1.3017
PIAFusion	0.07493	11.6930	0.0556	0.4531	1.2952
TarDAL	0.05820	12.8749	0.0354	0.4578	1.2759
SwinFusion	0.08005	11.4756	0.0528	0.4599	1.2940
CDDFuse	0.06888	12.1624	0.0432	0.4727	1.3108
NestFuse	0.06749	12.4418	0.0189	0.4674	1.3262
DIDFuse	0.05687	12.7122	0.0319	<u>0.4963</u>	1.2606
DeFusion	0.05569	<u>13.3174</u>	0.0061	0.4542	1.2943
CrossFuse	0.06520	12.4143	0.0180	0.4054	1.2082
PPIFuse	0.08873	11.1007	0.0434	0.4235	1.2758
TransFusion	0.05432	13.0251	<u>0.0074</u>	0.4759	1.3239
URFusion	0.06892	11.8386	0.0214	0.4839	<u>1.3290</u>
Ours	0.04388	13.9964	0.0078	0.5226	1.3711

4.4. Computational Efficiency

To provide a more comprehensive evaluation of computational efficiency, we report the average runtime, parameter count, and FLOPs of SiPFusion and the comparison methods. The input resolution was set to 640×480 for the computational evaluation.

As shown in Table 5, SiPFusion exhibits a moderate computational cost in terms of runtime, parameter count, and FLOPs. Although it is not among the most computationally efficient methods, this cost is associated with the progressive multi-scale feature extraction and Transformer-based global dependency modeling used to enhance local-detail preservation and global-context representation. We therefore acknowledge computational efficiency as a limitation of the current implementation. Future work will investigate lightweight attention mechanisms, efficient Transformer variants, token reduction, and model compression techniques to reduce inference latency while maintaining competitive fusion performance.

Table 5. Computational-efficiency comparison using an input resolution of 640×480 . Runtime is measured over 100 images. The best and second-best results are highlighted in **bold** and underlined, respectively. $\downarrow$ indicate that lower values are better.

Method	Runtime (s/100 Images) $\downarrow$	Params (M) $\downarrow$	FLOPs (G) $\downarrow$
MDLatLRR	<u>7.5729</u>	—	—
FusionDN	42.2988	1.1629	1031.4253
ReCoNet	8.4293	0.0036	6.4365
SeAFusion	69.4869	<u>0.1669</u>	<u>101.9904</u>
PIAFusion	10.4043	1.1761	721.9839
TarDAL	44.5793	0.2966	181.7248
SwinFusion	66.5644	0.9737	700.1487
CDDFuse	37.3328	1.1883	1104.9173
NestFuse	18.5203	2.7328	714.2179
DIDFuse	4.3965	0.2609	227.9080

Table 5. Cont.

Method	Runtime (s/100 Images) ↓	Params (M) ↓	FLOPs (G) ↓
DeFusion	176.1607	7.8745	141.1020
CrossFuse	88.1156	23.3102	271.9851
PPIFuse	20.3642	4.2485	155.5785
TransFusion	32.4340	53.1364	1822.6623
URFusion	19.8884	0.2433	126.7053
Ours	47.9313	18.0205	816.3377

4.5. Ablation Study

To investigate the contribution of each component in SiPFusion, we conducted an ablation study with eight settings. The ablation design (progressive strategy, feature extraction branches, fusion strategy, and loss terms) and the corresponding quantitative results are summarized in Table 6. In addition, Figure 7 provides qualitative comparisons (with local zoom-in regions) for all ablation settings to visually validate the reported effects. Specifically, the green boxes focus on the intensity/visibility of pedestrians (salient thermal targets), while the red boxes highlight thin-structure textures (e.g., tree-branch details), which together offer an intuitive validation of the quantitative trends.

Table 6. Ablation study of SiPFusion. The circled numbers denote different experimental settings. The checkmark and multiplication sign indicate that the corresponding component is included and excluded, respectively. “©” indicates that the SPF module is replaced with simple concatenation. The best and second-best results are highlighted in **bold** and underlined, respectively. ↑ and ↓ indicate that higher and lower values are better, respectively.

Setting	Ablation Design						Metrics				
	Progressive	MLFE	TGFE	SPF	L_{SRS}	L_{GSS}	MSE ↓	$PSNR$ ↑	$N^{AB/F}$ ↓	CC ↑	$MS-SSIM$ ↑
①	×	✓	✓	✓	✓	✓	0.11701	9.3823	0.0623	0.2262	0.8118
②	×	✓	×	✓	✓	×	0.06270	12.1141	0.0120	0.0458	1.2864
③	×	×	✓	✓	×	✓	0.10580	9.8144	0.0027	0.3163	0.9824
④	✓	✓	✓	©	×	✓	0.03823	14.4220	0.0793	0.1536	0.6932
⑤	✓	✓	✓	✓	×	✓	0.03088	<u>15.5145</u>	0.0125	0.6734	<u>1.4135</u>
⑥	✓	✓	✓	✓	✓	×	0.03200	15.4315	0.0103	<u>0.6743</u>	1.4061
⑦	✓	✓	✓	✓	×	×	0.03465	14.9193	0.0169	0.6703	1.3474
⑧	✓	✓	✓	✓	✓	✓	0.02925	15.8090	<u>0.0100</u>	0.6877	1.4203

In Table 6, “Progressive” refers to the organizational relationship between MLFE and TGFE. With Progressive enabled, the multi-scale local features produced by MLFE are used as the input to TGFE, forming the proposed serial MLFE–TGFE pathway. In Setting 1, Progressive is disabled while both MLFE and TGFE remain enabled. MLFE and TGFE therefore operate in parallel, with MLFE extracting local features and TGFE extracting global features independently before their outputs are provided to the subsequent fusion process. For Settings 2 and 3, only one feature-extraction module is retained, and the Progressive connection is therefore not applied.

The availability of the S-SIM terms depends on their corresponding intermediate features. L_{GSS} is calculated using the global features Φ_V^G and Φ_I^G produced by TGFE and is therefore unavailable when TGFE is removed in Setting 2. Similarly, L_{SRS} is calculated using the local selection masks $\widehat{\mathbf{PM}}_V^L$ and $\widehat{\mathbf{PM}}_I^L$ corresponding to the local features produced by MLFE and is therefore unavailable when MLFE is removed in Setting 3.

Effect of the progressive strategy. Comparing Setting 1 and Setting 8 (with all other components unchanged) shows that enabling the progressive local-to-global pipeline consistently improves fidelity- and structure-related metrics (lower MSE and higher

PSNR/MS-SSIM) as well as source consistency (higher *CC*). As shown in Figure 7, the progressive pipeline yields more reliable pedestrian intensity and target visibility in the green boxes while better preserving thin-structure textures (e.g., branch continuity and sharpness) in the red boxes, indicating that progressive cross-scale alignment and global dependency modeling are crucial for high-quality fusion.

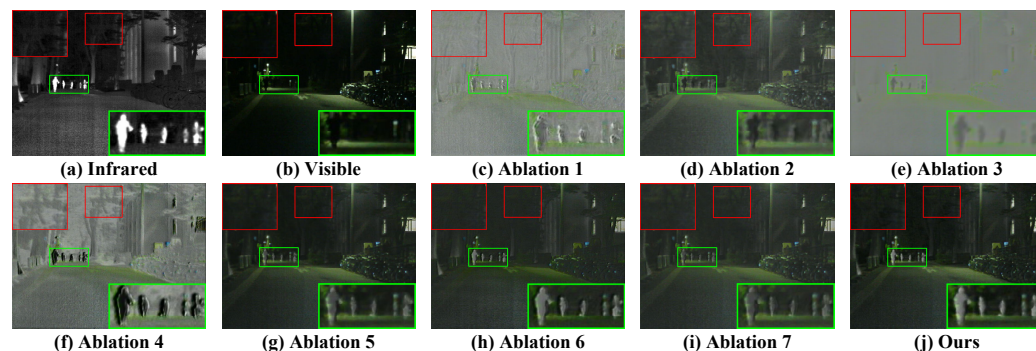


Figure 7. Qualitative comparisons of the ablation settings in Table 6. (a) Infrared image, (b) visible image, (c–i) correspond to Settings 1–7, and (j) is the full model (Setting 8). The local magnified regions are shown for clearer inspection of pedestrian intensity/visibility (green boxes) and fine texture details such as tree branches (red boxes).

Effect of the feature extraction branches (MLFE and TGFE). We evaluated MLFE by comparing Settings 3 and 5 and TGFE by comparing Settings 2 and 6. It should be noted that these matched comparisons are not strictly single-variable ablations because removing either MLFE or TGFE also disables the progressive connection between the two modules. Therefore, the results reflect the contribution of each feature-extraction module within the proposed progressive local-to-global pathway. In both matched comparisons, introducing the corresponding feature-extraction module produces clear improvements in *MSE*, *PSNR*, and *MS-SSIM*. Qualitatively, as shown in Figure 7, removing TGFE tends to cause less stable intensity organization for salient pedestrians in the green-boxed regions, leading to darker or lower-contrast targets and weakened source consistency. Removing MLFE often results in over-smoothing and loss of fine textures, which is particularly evident in the thin branches within the red-boxed regions. These observations suggest that MLFE and TGFE are complementary: MLFE preserves fine-grained local details, whereas TGFE contributes to global structural consistency.

Effect of the significance-preserving fusion module. To evaluate the fusion strategy, we compared the simple concatenation baseline (Setting 4) with the SPF-based fusion (Setting 7) with the same progressive encoder and without SRS/GSS. SPF yields consistent gains across all metrics, including better reconstruction fidelity (*MSE/PSNR*), fewer artifacts ($N^{AB/F}$), higher correlation (*CC*), and stronger structural similarity (*MS-SSIM*). The visual comparison in Figure 7 also shows that simple concatenation is more prone to suppressing salient-target intensity/visibility (green boxes) and attenuating thin-structure textures (red boxes), whereas SPF better preserves pedestrian intensity while retaining sharper branch details, validating the effectiveness of spatial selection.

Effect of SRS and GSS in S-SIM. We further disentangle the two terms in S-SIM by evaluating GSS-only (Setting 5), SRS-only (Setting 6), the model without S-SIM (Setting 7), and full S-SIM (Setting 8). Specifically, the contribution of GSS is evaluated by comparing Settings 6 and 8, which differ only in whether GSS is enabled. The contribution of SRS is evaluated by comparing Settings 5 and 8, which differ only in whether SRS is enabled. Comparing Settings 7 and 8 further evaluates the combined contribution of the two terms. Both single-term variants outperform the setting without S-SIM, while combining them

achieves the best overall performance across fidelity ($MSE/PSNR$), artifact suppression ($N^{AB/F}$), source consistency (CC), and structural similarity ($MS-SSIM$). In Figure 7, SRS-only tends to better preserve the intensity and visibility of salient pedestrians in the green-boxed regions, whereas GSS-only more effectively stabilizes structural consistency and reduces texture breakage on thin branches in the red boxes. Combining both terms simultaneously maintains salient-target intensity and fine texture fidelity, confirming the complementary roles of SRS and GSS in fusion.

To further examine the weighting strategy in GSS, we compared the proposed adaptive GSS weights with a fixed-weight variant. In the fixed-weight variant, the adaptive coefficients $\mathbf{W}_3$ and $\mathbf{W}_4$ were replaced with $\mathbf{W}_3 = \mathbf{W}_4 = 0.5$, while the remaining network architecture, loss function, training data, and optimization settings were kept unchanged. We also included the variant without GSS for reference. This comparison evaluated whether the global-feature-based adaptive weighting provides advantages over uniform weighting for the two source-specific structural similarity terms.

As shown in Table 7, the fixed-weight GSS variant improves MSE and $PSNR$ compared with removing GSS, but it does not consistently improve $N^{AB/F}$, CC , or $MS-SSIM$. In contrast, the proposed adaptive GSS weighting strategy achieves the best performance across all five metrics. These results indicate that using global feature responses to adjust the relative contributions of the visible- and infrared-specific structural constraints is more effective than assigning equal weights to the two terms.

Table 7. Ablation comparison of the GSS weighting strategy. The best and second-best results are highlighted in **bold** and underlined, respectively. $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively.

GSS Weighting Strategy	$MSE \downarrow$	$PSNR \uparrow$	$N^{AB/F} \downarrow$	$CC \uparrow$	$MS-SSIM \uparrow$
Without GSS	0.03200	15.4315	<u>0.0103</u>	<u>0.6743</u>	<u>1.4061</u>
Fixed weights ($\mathbf{W}_3 = \mathbf{W}_4 = 0.5$)	<u>0.03128</u>	<u>15.5396</u>	0.0108	0.6696	1.4033
Adaptive weights (Ours)	0.02925	15.8090	0.0100	0.6877	1.4203

In addition to the component ablation settings above, we further compared SPF with representative attention-style fusion modules. Specifically, we replaced SPF with CBAM-style [75] and NestFuse-style [34] fusion modules while keeping the remaining network architecture, loss function, training data, and optimization settings unchanged. The CBAM-style variant applies channel attention and spatial attention independently to each of the four feature branches and then sums the attended features. The NestFuse-style variant computes spatial and channel activity weights from the four feature branches and performs weighted fusion. These variants preserve the same input/output interface as SPF and only change the fusion rule. When the original S-SIM loss is used, the SRS masks are taken from the attention or weighting maps corresponding to the visible local and infrared local branches, while GSS is computed from the same TGFE global representations Φ_V^G and Φ_I^G .

As shown in Table 8, SPF achieves the best MSE , $PSNR$, and $MS-SSIM$ among the compared fusion modules, indicating stronger pixel-level fidelity and multi-scale structural preservation. Although the CBAM-style variant obtains a lower $N^{AB/F}$ and a higher CC , SPF maintains competitive artifact suppression and source correlation while improving fidelity and structural similarity. Compared with the NestFuse-style variant, SPF performs better across all evaluated metrics. These results indicate that the proposed branch-specific spatial selection strategy is more suitable for the four-branch local-global feature representation in SiPFusion than directly applying conventional attention-style feature refinement or activity-weighted fusion.

Table 8. Ablation comparison of different fusion modules. All variants use the same backbone, loss function, training data, and optimization settings. The best and second-best results are highlighted in **bold** and underlined, respectively. $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively.

Fusion Module	$MSE \downarrow$	$PSNR \uparrow$	$N^{AB/F} \downarrow$	$CC \uparrow$	$MS\text{-}SSIM \uparrow$
CBAM-style	<u>0.03140</u>	<u>15.5635</u>	0.0085	0.7080	<u>1.4177</u>
NestFuse-style	0.03491	15.0180	0.0209	0.6348	1.3670
SPF (Ours)	0.02925	15.8090	<u>0.0100</u>	<u>0.6877</u>	1.4203

5. Conclusions

This paper proposes a significance-preserving progressive fusion network for infrared and visible images. The algorithm first uses multi-scale blocks to extract multi-scale information from the source images and then inputs these multi-scale features into a Transformer-based feature extraction module to capture global dependency information. Subsequently, the SPF module adaptively maximizes the retention of useful information from the source images. Finally, the designed S-SIM function further strengthens global structural consistency while focusing on retaining effective information in salient regions. The proposed method is compared with 15 advanced algorithms on three datasets—MSRS, RoadScene, and TNO—and evaluated using five metrics: MSE , $PSNR$, $N^{AB/F}$, CC , and $MS\text{-}SSIM$. The proposed method achieves leading performance on most evaluated metrics and competitive overall results across the three datasets, although it does not consistently obtain the lowest $N^{AB/F}$ values. Overall, SiPFusion effectively preserves spatial structures and texture details while limiting visually apparent noise and artifacts, demonstrating its potential for infrared and visible image fusion. Despite the promising performance, our current implementation has a limitation in inference efficiency due to the progressive multi-scale encoder and the Transformer-based global modeling. In future work, we will investigate more efficient architectures to reduce latency, including lightweight attention and token reduction strategies, efficient Transformer variants, and model compression techniques such as pruning, quantization, and knowledge distillation, while maintaining high fusion quality. Future work will also investigate task-oriented evaluations, such as object detection and semantic segmentation, to further assess the practical impact of the fused images on downstream perception tasks.

6. Patents

The work reported in this manuscript has resulted in the following patent: Zhejiang University. A Method and System for Infrared and Visible Image Fusion Based on Global Integration of Multi-scale Information and Salient Region Preservation (in Chinese). Chinese Patent Application No. 202510249011.1, authorized on 27 October 2025.

Author Contributions: Conceptualization, J.L., X.L., S.C. and S.X.; methodology, J.L.; software, J.L.; validation, J.L., X.L., S.C. and S.X.; formal analysis, J.L.; investigation, J.L.; resources, X.L.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, X.L. and S.C.; visualization, J.L. and S.X.; supervision, X.L. and S.C.; project administration, X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under Grant No. 2024YFF1400900.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MSRS, RoadScene, and TNO datasets analyzed during this study are publicly available benchmarks. The data supporting the reported results can be found at their respective public repositories.

Acknowledgments: The authors sincerely appreciate the editors and anonymous reviewers for their valuable feedback, which greatly contributed to enhancing the quality of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Muller, A.C.; Narayanan, S. Cognitively-Engineered Multisensor Image Fusion for Military Applications. *Inf. Fusion* **2009**, *10*, 137–149.
2. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-Level Image Fusion: A Survey of the State of the Art. *Inf. Fusion* **2017**, *33*, 100–112.
3. Luo, Y.; Luo, Z. Infrared and Visible Image Fusion: Methods, Datasets, Applications, and Prospects. *Appl. Sci.* **2023**, *13*, 10891.
4. Tian, J.; Leng, Y.; Zhao, Z.; Xia, Y.; Sang, Y.; Hao, P.; Zhan, J.; Li, M.; Liu, H. Carbon Quantum Dots/Hydrogenated TiO₂ Nanobelt Heterostructures and Their Broad Spectrum Photocatalytic Properties under UV, Visible, and near-Infrared Irradiation. *Nano Energy* **2015**, *11*, 419–427.
5. Wang, X.; Zhang, C.; Hu, J.; Wen, Q.; Zhang, G.; Huang, M. AEFusion: Adaptive Enhanced Fusion of Visible and Infrared Images for Night Vision. *Remote Sens.* **2025**, *17*, 3129.
6. Yamamoto, N.; Saito, T.; Ogawa, S.; Ishimaru, I. Middle infrared (wavelength range: 8 μm –14 μm) 2-dimensional spectroscopy (total weight with electrical controller: 1.7 kg, total cost: Less than 10,000 USD) so-called hyper-spectral camera for unmanned air vehicles like drones. In *Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXII*, Baltimore, MD, USA, 18–21 April 2016; SPIE: Bellingham, WA, USA, 2016; Volume 9840, pp. 591–597.
7. Zhao, S.; Zhang, C.; Li, H.; Yue, M.; Shi, R. EMFusionNet: Attention-Guided Wavelet Mamba for Cross-Modal Fusion in Nighttime Infrared–Visible Images. *Infrared Phys. Technol.* **2026**, *153*, 106323.
8. Xie, W.; Pan, W.; Zhang, J.; Dai, S.; Xu, B.; Xu, C. DCD Fusion: A Dual-Consistency Decoupling Infrared and Visible Image Fusion Algorithm for Low-Light Scenes. *Infrared Phys. Technol.* **2026**, *152*, 106214.
9. Zhang, X.; Ye, P.; Leung, H.; Gong, K.; Xiao, G. Object Fusion Tracking Based on Visible and Infrared Images: A Comprehensive Review. *Inf. Fusion* **2020**, *63*, 166–187.
10. Kumar, P.; Mittal, A.; Kumar, P. Fusion of Thermal Infrared and Visible Spectrum Video for Robust Surveillance. In *Proceedings of the Computer Vision, Graphics and Image Processing*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 528–539.
11. Qin, X.; Wang, E.; Zhou, S.; Wang, B.; Kasabov, N.K. TSPFusion: Text-Guided Semantic Perception for Infrared and Visible Image Fusion. *Infrared Phys. Technol.* **2026**, *153*, 106324.
12. Zhang, Y.; Li, S.; Fang, A.; He, X.; Zhu, D.; Wu, K.; Xiao, X. KANFuse: Enhancing Infrared and Visible Image Fusion through Nonlinear Representation Modeling. *Infrared Phys. Technol.* **2026**, *153*, 106319.
13. Xiang, X.; Zhou, G.; Niu, B.; Pan, Z.; Huang, L.; Li, W.; Wen, Z.; Qi, J.; Gao, W. Infrared-Visible Image Fusion Meets Object Detection: Towards Unified Optimization for Multimodal Perception. *Remote Sens.* **2025**, *17*, 3637.
14. Zhuo, Z.; Lu, R.; Yao, Y.; Wang, S.; Zheng, Z.; Zhang, J.; Yang, X. TAF-YOLO: A Small-Object Detection Network for UAV Aerial Imagery via Visible and Infrared Adaptive Fusion. *Remote Sens.* **2025**, *17*, 3936.
15. Chen, J.; Li, X.; Luo, L.; Mei, X.; Ma, J. Infrared and Visible Image Fusion Based on Target-Enhanced Multiscale Transform Decomposition. *Inf. Sci.* **2020**, *508*, 64–78.
16. Li, H.; Wu, X.J.; Kittler, J. MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746.
17. Liu, Y.; Chen, X.; Ward, R.K.; Jane Wang, Z. Image Fusion with Convolutional Sparse Representation. *IEEE Signal Process. Lett.* **2016**, *23*, 1882–1886.
18. Zhang, Q.; Liu, Y.; Blum, R.S.; Han, J.; Tao, D. Sparse Representation Based Multi-Sensor Image Fusion for Multi-Focus and Multi-Modality Images: A Review. *Inf. Fusion* **2018**, *40*, 57–75.
19. Mou, J.; Gao, W.; Song, Z. Image Fusion Based on Non-Negative Matrix Factorization and Infrared Feature Extraction. In *Proceedings of the 2013 6th International Congress on Image and Signal Processing (CISP)*, Hangzhou, China, 16–18 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1046–1050.
20. Fu, Z.; Wang, X.; Xu, J.; Zhou, N.; Zhao, Y. Infrared and Visible Images Fusion Based on RPCA and NSCT. *Infrared Phys. Technol.* **2016**, *77*, 114–123.
21. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and Visible Image Fusion Based on Visual Saliency Map and Weighted Least Square Optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17.
22. Jun, C.; Lei, C.; Wei, L.; Yang, Y. Fusion of Near-Infrared and Visible Images Based on Saliency-Map-Guided Multi-Scale Transformation Decomposition. *Multimed. Tools Appl.* **2023**, *82*, 34631–34651.

23. Sun, G.; Kong, W.; Yu, L.; Li, T.; Zhou, K. A Novel Infrared and Visible Image Fusion Method via Adaptive Offsets Enhanced Hybrid Attention. *Infrared Phys. Technol.* **2026**, *152*, 106269.
24. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. FusionDN: A Unified Densely Connected Network for Image Fusion. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12484–12491.
25. Song, W.; Gao, M.; Li, Q.; Guo, X.; Wang, Z.; Jeon, G. Optimizing Nighttime Infrared and Visible Image Fusion for Long-Haul Tactile Internet. *IEEE Trans. Consum. Electron.* **2024**, *70*, 4277–4286.
26. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A Progressive Infrared and Visible Image Fusion Network Based on Illumination Aware. *Inf. Fusion* **2022**, *83–84*, 79–92.
27. Tang, L.; Yuan, J.; Ma, J. Image Fusion in the Loop of High-Level Vision Tasks: A Semantic-Aware Real-Time Infrared and Visible Image Fusion Network. *Inf. Fusion* **2022**, *82*, 28–42.
28. Ma, J.; Zhang, H.; Shao, Z.; Liang, P.; Xu, H. GANMcC: A Generative Adversarial Network With Multiclassification Constraints for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5005014.
29. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.P. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995.
30. Xiao, Q.; Jin, H.; Su, H.; Zuo, F.; Zhang, Y.; Xiao, Z.; Wang, B. SPGFusion: A Semantic Prior Guided Infrared and Visible Image Fusion Network. In *Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 2405–2409.
31. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 5792–5801.
32. Xiao, Q.; Jin, H.; Su, H.; Zhang, Y.; Xiao, Z.; Wang, B. SPDFusion: A Semantic Prior Knowledge-Driven Method for Infrared and Visible Image Fusion. *IEEE Trans. Multimed.* **2025**, *27*, 1691–1705.
33. Li, H.; Wu, X.J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623.
34. Li, H.; Wu, X.J.; Durrani, T. NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656.
35. Li, H.; Wu, X.J.; Kittler, J. RFN-Nest: An End-to-End Residual Fusion Network for Infrared and Visible Images. *Inf. Fusion* **2021**, *73*, 72–86.
36. Wang, Z.; Wu, Y.; Wang, J.; Xu, J.; Shao, W. Res2Fusion: Infrared and Visible Image Fusion Based on Dense Res2net and Double Nonlocal Attention Models. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5005012.
37. Karim, S.; Tong, G.; Li, J.; Yu, X.; Hao, J.; Qadir, A.; Yu, Y. MTDFusion: A Multilayer Triple Dense Network for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2023**, *73*, 5010117.
38. Zhao, Z.; Xu, S.; Zhang, C.; Liu, J.; Li, P.; Zhang, J. DIDFuse: Deep Image Decomposition for Infrared and Visible Image Fusion. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021*; pp. 970–976.
39. Xu, H.; Wang, X.; Ma, J. DRF: Disentangled Representation for Visible and Infrared Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5006713.
40. Yang, Y.; Zhou, N.; Wan, W.; Huang, S. MACCNet: Multiscale Attention and Cross-Convolutional Network for Infrared and Visible Image Fusion. *IEEE Sens. J.* **2024**, *24*, 16587–16600.
41. Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; Van Gool, L. CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 5906–5916.
42. Liu, Q.; Pi, J.; Gao, P.; Yuan, D. STFNet: Self-Supervised Transformer for Infrared and Visible Image Fusion. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 1513–1526.
43. Xu, P.; Luo, G.; Liu, J.; Zhou, X.; Song, D. DIFuse: Dual-Domain Interactive Fusion of Cross-Modality Infrared and Visible Images with Decomposition Deep Network. *Infrared Phys. Technol.* **2025**, *153*, 106342.
44. Zhao, Z.; Ma, Y.; Huang, J.; Wu, K.; Wang, G.; Fan, F. CrossMamba: Cross-Modal Features Mixing via Mamba for Infrared and Visible Image Fusion. *Infrared Phys. Technol.* **2026**, *152*, 106234.
45. Chen, J.; Zhu, T.; Wang, G. Target-Aware Fusion: A Diffusion Model for Infrared and Visible Image Integration to Enhance Object Detection. *Remote Sens.* **2026**, *18*, 1545.
46. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518.
47. Zhao, Y.; Zheng, Q.; Zhu, P.; Zhang, X.; Ma, W. TUFusion: A Transformer-Based Universal Fusion Algorithm for Multimodal Images. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 1712–1725.

48. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. CMT: Convolutional Neural Networks Meet Vision Transformers. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 12165–12175.
49. Huang, Y.; Liu, Z.; Lu, H.; Wang, W.; Lan, R. Brighten up Images via Dual-Branch Structure-Texture Awareness Feature Interaction. *IEEE Signal Process. Lett.* **2024**, *31*, 46–50.
50. Rao, D.; Xu, T.; Wu, X.J. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. *IEEE Trans. Image Process.* **2022**, *early access*.
51. Liu, X.; Gao, H.; Miao, Q.; Xi, Y.; Ai, Y.; Gao, D. MFST: Multi-Modal Feature Self-Adaptive Transformer for Infrared and Visible Image Fusion. *Remote Sens.* **2022**, *14*, 3233.
52. Tang, W.; He, F.; Liu, Y.; Duan, Y. MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer. *IEEE Trans. Image Process.* **2022**, *31*, 5134–5149.
53. Li, J.; Zhu, J.; Li, C.; Chen, X.; Yang, B. CGTF: Convolution-Guided Transformer for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5012314.
54. Hu, C.; Hu, Z.; Xu, C.; Song, Z.; Mei, L. Multi-Scale Frequency-Aware Representation Learning for Infrared and Visible Image Fusion. *Remote Sens.* **2026**, *18*, 1178.
55. Huang, S.; Kong, X.; Yang, Y.; Wan, W.; Song, Z. FTSFN: A Two-Stage Feature Transfer and Supplement Fusion Network for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5006715.
56. Pan, C.; Jiang, Q.; Zheng, H.; Wang, P.; Jin, X.; Yao, S.; Zhou, W. DANet: A Dual-Branch Framework with Diffusion-Integrated Autoencoder for Infrared-Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5031713.
57. Chen, J.; Ding, J.; Yu, Y.; Gong, W. THFuse: An Infrared and Visible Image Fusion Network Using Transformer and Hybrid Feature Extractor. *Neurocomputing* **2023**, *527*, 71–82.
58. Tang, W.; He, F.; Liu, Y.; Duan, Y.; Si, T. DATFuse: Infrared and Visible Image Fusion via Dual Attention Transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3159–3172.
59. Tang, W.; He, F.; Liu, Y. YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer. *IEEE Trans. Multimed.* **2023**, *25*, 5413–5428.
60. Qu, L.; Liu, S.; Wang, M.; Li, S.; Yin, S.; Song, Z. Trans2Fuse: Empowering Image Fusion through Self-Supervised Learning and Multi-Modal Transformations via Transformer Networks. *Expert Syst. Appl.* **2024**, *236*, 121363.
61. Yan, H.; Xiong, S.; Wang, L.; Jian, L.; Vivone, G. ATFusion: An Alternate Cross-Attention Transformer Network for Infrared and Visible Image Fusion. *Infrared Phys. Technol.* **2026**, *152*, 106252.
62. Zhang, Q.; Liu, Q.; Yuan, D.; Liu, Y. One-Stream One-Stage Pure Transformer for Infrared and Visible Image Fusion. *Infrared Phys. Technol.* **2025**, *151*, 106067.
63. Fan, Y.; Ran, W.; Tan, K.; Liu, Q.; Yuan, D.; Li, X.; Liu, Y. D2Fusion: Dual-Domain Feature Decoupling for Infrared and Visible Image Fusion. *Knowl.-Based Syst.* **2026**, *332*, 114892.
64. He, J.; Cheng, J.; Liu, T.; Cheng, B.; Pan, X.; Cai, Y. Mamba-Based Infrared and Visible Images Fusion Method. *Remote Sens.* **2026**, *18*, 636.
65. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-Domain Long-Range Learning for General Image Fusion via Swin Transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217.
66. Li, H.; Wu, X.J. CrossFuse: A Novel Cross Attention Mechanism Based Infrared and Visible Image Fusion Approach. *Inf. Fusion* **2024**, *103*, 102147.
67. Zhang, Q.; Liu, Q.; Yuan, D.; Li, X.; Liu, Y. PPIFuse: Physical Priors Injected Infrared and Visible Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *36*, 4521–4535.
68. Qian, Y.; An, R.; Liu, G.; Tang, H.; Xiao, G.; Bavirisetti, D.P. TransFusion: Transfer Learning-Driven Adaptive Fusion Network for Infrared and Visible Image. *Infrared Phys. Technol.* **2025**, *150*, 105906.
69. Xu, H.; Yi, X.; Lu, C.; Liu, G.; Ma, J. URFusion: Unsupervised Unified Degradation-Robust Image Fusion Network. *IEEE Trans. Image Process.* **2025**, *34*, 5803–5818.
70. Toet, A.; Hogervorst, M.A. Progress in color night vision. *Opt. Eng.* **2012**, *51*, 010901.
71. Tang, L.; Zhang, H.; Xu, H.; Ma, J. Deep learning-based image fusion: A survey. *J. Image Graph.* **2023**, *28*, 3–36.
72. Wang, Z.; Simoncelli, E.; Bovik, A. Multiscale Structural Similarity for Image Quality Assessment. In *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003*; IEEE: Piscataway, NJ, USA, 2003; pp. 1398–1402.
73. Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; Luo, Z. ReCoNet: Recurrent Correction Network for Fast and Efficient Multi-Modality Image Fusion. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; Volume 13678, pp. 539–555.

74. Liang, P.; Jiang, J.; Liu, X.; Ma, J. Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; Volume 13678, pp. 719–735.
75. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision – ECCV 2018*; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.