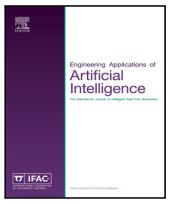
Engineering Applications of Artificial Intelligence 159 (2025) 111728

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

Multi-scale frequency attention fusion network for infrared and visible image fusion

Yong Wang [a], Xueyuan Zhao [a,*], Jianfei Pu [a], Lulu Zhang [a], Duoqian Miao [b]

[a] School of Artificial Intelligence, Chongqing University of Technology, Chongqing, 401135, China
[b] School of Computer Science and Technology, Tongji University, Shanghai, 201804, China

ARTICLE INFO

Keywords:
Image fusion
Multi-scale convolution
Attention mechanism
Frequency domain

ABSTRACT

The goal of visible and infrared image fusion is to generate a composite image that not only preserves fine-grained textures from visible images but also highlights salient targets from infrared images. However, existing methods often struggle to capture detailed features and typically rely solely on spatial domain information, overlooking the complementary advantages offered by frequency domain features. To address these limitations, we propose a feature-level fusion approach based on a multi-scale frequency attention fusion network, which incorporates a spatial-frequency attention fusion module with cross-attention and a multi-scale depthwise separable convolution block equipped with coordinate attention. Moreover, a multi-scale compensation fusion is incorporated within the spatial frequency attention fusion module to reduce cross-modal domain discrepancies. This work applies deep learning-based artificial intelligence (AI) techniques to the field of multi-modal image fusion. Experiments on three public datasets show that our method achieves competitive performance across multiple evaluation metrics compared to state-of-the-art methods. Our method demonstrates superior performance in preserving fine details and enhancing target saliency. The code will be available at https://github.com/ioschunsheng1230.

1. Introduction

Due to the physical limitations of imaging sensors, optical imaging principles, and viewing angles, a single image sensor often fails to extract sufficient information from a scene to fully represent its semantic and structural content, resulting in incomplete or limited imagery. Infrared sensors capture thermal radiation emitted from objects in the scene, enabling the detection of targets such as humans or vehicles even under challenging conditions such as low illumination, occlusion, or concealment. However, infrared images typically suffer from low spatial resolution and a lack of fine-grained texture details. In contrast, visible sensors capture reflected light from objects or environments, producing images with high spatial resolution and rich texture information, which are well-suited for human visual perception. Nevertheless, visible images are highly susceptible to environmental factors such as nighttime darkness, fog, or occlusion. As illustrated in Fig. 1, visible images at night may offer broader environmental context but tend to obscure pedestrians due to intense lighting or low contrast. Thanks to the distinct imaging mechanism of infrared devices, infrared images can clearly depict pedestrians, as shown in Fig. 1(d). In low-light conditions, the fused image effectively retains both prominent target features and nighttime environmental texture, thereby enhancing both human visual understanding and machine vision performance. Recent work has also explored the enhancement of infrared imagery under complex scenes, such as low-light or nighttime environments, to better extract salient features and improve perceptual quality (Maruschak et al., 2024). However, these methods often focus on enhancing a single modality, whereas multi-modal fusion provides a more comprehensive representation by leveraging complementary advantages from both visible and infrared sources. The complementary characteristics of different sensors and the potential to produce visually informative results make image fusion widely applicable in practical scenarios such as nighttime driving assistance, video surveillance (Paramanandham and Rajendiran, 2018), object detection (Jain et al., 2023), tracking (Zhang et al., 2021a), and semantic segmentation (Zhang et al., 2021b).

Image fusion technology that combines visible and infrared light has become more popular in recent years because of its usefulness. Current There are two types of fusion approaches: deep learning-based techniques and traditional procedures (Ma et al., 2019b):

(1) Traditional methods primarily rely on mathematical transformations and typically consist of three stages: feature extraction, feature

* Corresponding author.
E-mail addresses: ywang@cqut.edu.cn (Y. Wang), zxy19991230lntl@163.com (X. Zhao), pjf674786205@163.com (J. Pu), 1286852139@qq.com (L. Zhang), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.engappai.2025.111728
Received 12 February 2025; Received in revised form 23 June 2025; Accepted 7 July 2025
Available online 19 July 2025

Fig. 1. The comparison of fusion results between CrossFuse and our method under low-light conditions.

fusion, and feature reconstruction. Features are first extracted from source images using specific transformations, then fused according to predefined rules, and finally reconstructed into a fused image via inverse transformation. Depending on the mathematical tools used, traditional methods can be further classified into multi-scale decomposition-based methods (Li et al., 2020b; Liu et al., 2014), subspace clustering-based methods (Cvejic et al., 2007), sparse representation-based methods (Liu et al., 2016), and optimization-based strategies (Ma et al., 2016; Gao et al., 2024). However, these methods heavily depend on hand-crafted features and fusion rules, resulting in limited generalization ability and difficulty in adapting to complex scenarios. Additionally, their performance is often inadequate when dealing with large-scale modality differences and nonlinear feature coupling.

(2) The rise of deep learning has significantly advanced (Gao et al., 2024) the development of image fusion techniques. Current deep learning-based fusion methods mainly fall into three categories: autoencoder (AE)-based methods (Li and Wu, 2018; Li et al., 2020a), convolutional neural network (CNN)-based methods (Xu et al., 2020; Ma et al., 2021; Zhang and Ma, 2021), and generative adversarial network (GAN)-based methods (Ma et al., 2019a; Liu et al., 2022) . AE-based methods leverage pre-training to obtain robust feature extraction and reconstruction capabilities, combining them with fusion strategies to generate fused images. CNN-based approaches typically adopt end-to-end architectures for automatic feature learning and fusion modeling. GAN-based methods utilize adversarial training between a generator and a discriminator to produce visually realistic fusion results. Although these methods demonstrate strong nonlinear modeling abilities and high fusion quality, they also face several limitations. Firstly, most of them focus on spatial domain feature modeling and overlook critical structural and detail information embedded in the frequency domain. Secondly, models such as CNNs are constrained by their local receptive fields, making it difficult to capture global context and long-range dependencies. This often leads to inadequate representation of salient targets and loss of texture details in cross-modal fusion scenarios. To address these issues, recent studies have introduced Transformer (Vaswani et al., 2017) architectures into image fusion tasks, leveraging their powerful self-attention mechanisms to globally model and integrate complementary features from infrared and visible images, thereby enhancing both detail preservation and target saliency (Chang et al., 2023; Zhou et al., 2023). For instance, SwinFusion (Ma et al., 2022) proposes combining intra-domain self-attention with inter-domain cross-attention to achieve efficient coordination and enhancement of multimodal information. Nevertheless, existing approaches still primarily focus on spatial domain fusion, with insufficient modeling and utilization of frequency domain information—this remains a critical bottleneck for further improving fusion quality.

To address the aforementioned challenges, we propose the multi-scale frequency attention fusion network (MSFAFusion), a novel infrared and visible image fusion network based on feature-level fusion. Specifically, our network targets two key limitations in existing methods: (1) insufficient modeling capability for image detail hierarchies during the feature extraction stage and (2) the underutilization of frequency domain formation. To tackle these issues, we design a multi-scale depthwise separable convolution block (Block) with coordinate attention (Hou et al., 2021) . This module employs convolution kernels of varying sizes (Sun et al., 2025; Szegedy et al., 2016) to simulate multiple receptive fields, enabling the extraction of multi-scale shallow features from the source images. The architecture helps capture both local details and global structural information, enhancing the network's ability to perceive edge details and salient targets. In the field of computer vision, multi-scale feature extraction and receptive-field attention mechanisms have been widely adopted and demonstrated remarkable effectiveness, particularly in small object detection tasks. For example, they (Tao et al., 2024) proposed an enhanced feature extraction method based on receptive-field attention and multi-scale feature fusion, which significantly improved the performance of industrial small object detection. Inspired by this work, our study incorporates coordinate attention mechanisms to further strengthen the feature module's capability in capturing fine-grained details and structural information from infrared and visible images, thereby enhancing the detail quality and structural integrity of the fused images. To further exploit the potential of frequency domain information in image fusion, we design the spatial frequency attention fusion module (SFAFM), which combines cross-attention mechanisms (Lin et al., 2022) with frequency domain (Khayam, 2003) feature modeling. This module consists of three components: cross-attention spatial fusion (CASF), frequency domain fusion (FDF), and multi-scale compensation fusion (MCF). CASF and FDF are responsible for generating fused features in the spatial and frequency domains, respectively, enabling joint modeling of multi-domain information. It is worth emphasizing that the cross-attention mechanism not only facilitates effective interaction between modalities but also significantly enhances the response to salient targets in infrared images, improving the model's sensitivity to key regions. Building on this, the MCF module performs multi-scale modeling and feature compensation on the fused features from both spatial and frequency domains, effectively mitigating the cross-domain discrepancies. Given that MCF is introduced to align deep features across different modalities, we also incorporate cross-attention mechanisms here to further strengthen inter-domain adaptation and fusion. The overall network adopts a three-branch architecture, allowing feature extraction and fusion to be carried out in parallel and dynamically synchronized. This design not only improves the expressive power of the features but also significantly enhances computational efficiency and fusion quality. Experiments on three public datasets show that our method achieves competitive performance across multiple evaluation metrics compared to state-of-the-art methods. Evaluation on three public datasets confirms that our method achieves strong and competitive performance across diverse metrics relative to existing state-of-the-art algorithms. Our method outperforms the latest approaches on three public datasets in terms of the evaluation metrics AG, Qabf, and SF. For example, on the MSRS dataset, the corresponding values for MaeFuse (Li et al., 2025) are 3.461, 0.502, and 9.569; for MLFuse (Lei et al., 2025), they are 3.439, 0.525, and 11.551; while our method achieves 4.003, 0.676, and 12.280, respectively. These results clearly demonstrate the effectiveness of our proposed method.

This paper's primary contributions can be summed up as follows:

- We propose multi-scale frequency attention fusion network (MSFAFusion), a feature-level infrared and visible image fusion network that effectively integrates spatial and frequency domain information for deep cross-domain feature fusion.
- We design a multi-scale depthwise separable convolution block (MDSCBlock) integrated with a coordinate attention mechanism to enhance the network's capability in capturing salient target features and preserving fine-grained texture details.
- We propose the spatial frequency attention fusion module (SFAFM), which combines cross-attention and frequency-domain modeling. It enables collaborative fusion of spatial and frequency features and uses multi-scale compensation to reduce domain discrepancies, enhancing detail preservation and structural consistency in the fused image.
- We conducted experiments using existing datasets and performed both qualitative and quantitative comparisons between our algorithm and state-of-the-art methods. The results indicate that our approach offers advantages in preserving the desired saliency and texture features.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed method in detail, including the adopted frequency-domain techniques, overall network architecture, module designs, and loss functions. Section 4 provides the experimental settings, implementation details, comparison results, and performance on downstream tasks. Finally, Section 5 concludes the paper and discusses future work.

2. Related work

This section will provide a brief overview of the current state of image fusion techniques, including both deep learning-based and traditional image fusion methods. Additionally, we will delve into the importance of frequency-domain transformation in computer vision tasks. As shown in Table 1, for ease of reading and comparison, we have summarized the advantages and disadvantages of each fusion method according to their respective categories.

2.1. Infrared and visible image fusion

The goal of fusing visible and infrared images is to create a composite image that preserves rich texture features while emphasizing the important elements from the original images. Traditional image fusion algorithms typically operate in the field of transformation or space, where mathematical operations convert images into the transform domain, followed by activity level measurement and manually developed fusion rules to accomplish the fusion of images (Ma et al., 2019b). Multi-scale transformation-based fusion frameworks are the main component of classical traditional image fusion frameworks (Zhang et al., 2020b). Image fusion techniques based on multi-scale transformations usually start with multi-scale decomposition on two source images, then fuse the high-frequency and low-frequency sub-bands using different fusion criteria at each scale, and finally apply multi-scale inverse transformation to create the fused image. For instance, MDLatLRR (Li et al., 2020b) is a multi-level image decomposition technique based on latent low-rank representation, which achieves superior fusion performance in evaluations by breaking down source images into detailed parts and base parts and using nuclear-norm and averaging strategies, respectively. Pyramid-based fusion is a common multi-scale transformation technique (Toet, 1989), wavelet-based fusion (Zhang, 2010). In addition, there are saliency-based fusion frameworks (Jerripothula et al., 2022) and variational model-based fusion frameworks (Ning et al., 2013). Even though conventional image fusion algorithms can typically provide fused images that are reasonably adequate, there are still certain problems that prevent them from being developed further. Firstly, existing methods often adopt the same transformations or representations to extract features from source images but fail to fully consider the intrinsic differences between multi-source images. In other words, these methods overlook the inherent differences in characteristics and information distribution among source images. Second, in an effort to improve fusion performance, the complexity of manually generated activity level measurements and fusion rules has increased, making it difficult for them to adjust to complex fusion scenarios (Ma et al., 2023).

Due to the exceptional feature learning skills of neural networks, deep learning has become widely used for a variety of applications and has been acknowledged by researchers for its potent learning powers (Akhtar et al., 2025). Following widespread application in the image processing field, numerous innovative algorithms combining deep learning with infrared and visible image fusion (IVIF) tasks have emerged. End-to-end fusion frameworks and pre-trained fusion frameworks are two broad categories into which deep learning-based methods for fusing visible and infrared images can be divided. Autoencoders (AE), also known as pre-trained fusion frameworks, typically comprise two primary phases. In order to extract features and reconstruct images, an autoencoder is first pre-trained on a sizable dataset. The deep features that have been retrieved from several source images are then integrated using a manually created fusion method to accomplish image fusion. Li et al. proposed the groundbreaking pre-trained fusion model DenseFuse (Li and Wu, 2018), which consists of three primary components: an encoder layer, a fusion layer, and a decoder layer. By utilizing the properties of densely linked networks, features are retrieved in the encoder layer. By using skip connections, information from both shallow and deep features is successfully preserved, producing fused images with superior visual quality. In a similar vein, DIVFusion (Tang et al., 2023b) used several encoders to separate the illumination and reflectance components of visible light images, which enhanced fusion performance in low light, thus expanding upon the Retinex theory.

However, these manually designed fusion strategies are not always suitable for deep features, which limits the performance of AE-based fusion frameworks. Techniques that make use of Convolutional Neural Networks (CNN) carry over the fundamental concepts of conventional optimization techniques and have been extensively used because of their exceptional feature extraction abilities in a range of image fusion problems. By creating network designs and loss functions, without the trouble of manually establishing fusion rules, CNN-based image fusion frameworks provide end-to-end feature extraction, feature fusion, and image reconstruction (Ma et al., 2021). The CNN-based fusion architecture guides the network for end-to-end training by building loss functions according to the way the fused image resembles the original images (Han et al., 2022). PIAFusion (Tang et al., 2022) proposed a progressive image fusion network based on illumination awareness. By designing an illumination-aware subnetwork to estimate illumination distribution and compute illumination probabilities, the method adaptively perceives intensity variations. Structural Similarity Index (SSIM) loss (Long et al., 2021) and perceptual loss (Ma et al., 2020a) are introduced to construct the loss function, effectively guiding the training process. This approach enables efficient cross-modal information fusion and enhances the robustness of the fusion results under varying illumination conditions. In addition, unsupervised networks such as the dense residual-based RXDNFuse (Long et al., 2021), the general image fusion framework IFCNN (Zhang et al., 2020a), the SDNet (Zhang and Ma, 2021) focusing on both decomposition and fusion stages, and the unified multi-task fusion model U2Fusion (Xu et al., 2020), which considers the correlations among various image fusion tasks, have also been developed.

However, the lack of real fused images to use as a reference has led researchers to incorporate Generative Adversarial Networks (GAN) into the learning paradigm. With FusionGAN (Ma et al., 2019a), the generator is forced to maintain more texture features from visible images by use of a discriminator. The use of GANs in image fusion

Table 1
Summary of advantages and limitations of some popular image fusion methods.

Method	Advantages	Limitation
CSR (Liu et al., 2016)	This method effectively extracts structural features from multi-source images using convolutional sparse representation, enhancing detail preservation in the fused image.	It relies on manually designed sparse dictionaries and parameter settings, making it less adaptable to complex and diverse fusion scenarios.
MDLatLRR (Li et al., 2020b)	It utilizes a novel multi-level decomposition based on latent low-rank representation to effectively separate and fuse salient targets and background information from different modalities.	The method involves handcrafted decomposition processes and lacks end-to-end optimization, limiting its adaptability and scalability in more complex fusion tasks.
DenseFuse (Li and Wu, 2018)	End-to-end training with a lightweight architecture that effectively preserves details and target information.	The fusion strategy is simple and primarily relies on spatial-domain features, making it difficult to capture global and frequency-domain information.
DIVFusion (Tang et al., 2023b)	By separating structural and detail information and employing multi-scale encoding, the method achieves richer preservation of details and structures.	The model is complex and computationally expensive, with limited utilization of frequency-domain information and high training requirements.
SDNet (Zhang and Ma, 2021)	Utilizes a decoupled network structure to effectively separate features from different modalities, enhancing the representational capacity of fusion.	The structure is relatively complex and difficult to train, and the preservation of fine details still has room for improvement.
U2Fusion (Xu et al., 2020)	Employs a dual-encoder architecture to enhance multi-scale feature extraction, thereby improving the quality of fused images.	The model has a large number of parameters and high computational cost, limiting its real-time performance.
FusionGAN (Ma et al., 2019a)	Introduces a generative adversarial mechanism, where the discriminator guides the generator to produce more realistic and consistent fused images.	The adversarial training process is unstable, prone to mode collapse, and requires careful hyperparameter tuning.
GANMcC (Ma et al., 2020b)	Leverages multi-channel attention mechanisms and collaborative adversarial training to effectively enhance the fusion of multimodal features and preserve salient targets.	The network structure is complex, with long training times and high hardware requirements, making deployment less friendly.
TCCFusion (Tang et al., 2023a)	Based on transformer and cross-correlation mechanisms, it effectively captures deep semantic relationships across spatial and modal dimensions.	The model demands high computational resources, and its ability to preserve fine texture details remains limited.
CrossFuse (Li and Wu, 2024)	Introduces a cross-attention mechanism focusing on modality complementarity, enhancing the collaborative fusion of infrared and visible features.	The network structure is relatively complex, and the two-stage training process increases implementation and optimization difficulty.

was initially introduced by Ma et al. Many traditional GAN-based techniques, such as GANMcC (Ma et al., 2020b) and SDDGAN (Zhou et al., 2021b), have surfaced since FusionGAN. Additionally, TarDAL (Liu et al., 2022) was designed to integrate image fusion with object detection tasks.

2.2. Attention mechanism in image fusion

Convolutional neural networks (CNNs) and their variants have been widely applied in the image fusion field because of their superior generalization performance and powerful feature extraction capabilities. Moreover, the network structures of CNNs are relatively mature, and corresponding hardware accelerators can significantly enhance their computational efficiency. However, CNNs also have some notable limitations. They struggle to capture dependencies between distant components of a picture and are inefficient at extracting global information due to their inherently small receptive fields. These issues can have a detrimental effect on the outcomes of picture fusion. Sadly, CNNs are used as feature extractors in nearly all image fusion frameworks now in use, but they are unable to create long-range correlations within the images (Zhou et al., 2021a). The selective attention process in human vision and cognition has been identified by neuroscientists, which enables people to choose to concentrate on the intriguing aspects of complicated data for additional processing and examination. As a result, neural network architectures like the attention mechanism have been frequently employed as the basis for Transformer (Vaswani et al., 2017). With the continuous development of attention mechanisms, they have also been progressively added to the image fusion field. Recent transformer-based designs have been introduced into the image fusion area via advances, leveraging their powerful global modeling capabilities. TCCFusion (Tang et al., 2023a), a transformer and cross-correlation-based methodology for fusing visible and infrared images to demonstrate better feature alignment between modalities. Similarly, AFT (Chang et al., 2023) was proposed as a flexible Transformer-based framework that dynamically fuses features for infrared and visible image fusion, enabling more adaptive and effective fusion strategies. Further pushing the limits of infrared and visible picture fusion, SCGRFuse (Wang et al., 2024) combines gradient aggregation with residual dense blocks and spatial/channel attention processes. These works highlight the increasing relevance of transformer-based and attention-enhanced approaches in modern image fusion tasks. CrossFuse (Li and Wu, 2024) further integrates the cross-attention mechanism into the image fusion network. The cross-attention mechanism effectively captures the correlation and complementarity between modalities by comparing and dynamically focusing on the features of the two modalities. Specifically, when fusing visible and infrared images, visible images often have complex texture features, while infrared images usually emphasize important target information. The cross-attention mechanism can adaptively enhance the feature interaction between the two, effectively preserving salient targets and detailed textures, thereby generating higher-quality fused images.

2.3. Frequency domain transform learning

In signal processing, frequency analysis has long been a potent instrument. Recent years have seen the emergence of certain applications that bring frequency analysis into the deep learning space. In (Ehrlich and Davis, 2019; Gueguen et al., 2018), frequency analysis was incorporated into CNNs through JPEG encoding. However, most existing image fusion methods are primarily conducted in the spatial domain, often overlooking the critical information embedded in the frequency domain. Recently, some frequency-based methods have been proposed, such as FISCNet (Zheng et al., 2024) and SFINet (Zhou et al., 2025), which leverage Fourier transform (Yu et al., 2022) to enhance image fusion performance. The Fourier Transform can effectively

capture global frequency information but has limited performance in capturing local features and is prone to boundary artifacts. In addition, frequency analysis methods include wavelet transform (Sifuzzaman et al., 2009) and discrete cosine transform (DCT) (Khayam, 2003). Wavelet transform excels in multi-scale analysis and local feature extraction but may introduce redundant information due to its complex computation process. In contrast, DCT efficiently concentrates energy into low-frequency components, preserves critical edges and texture information, reduces redundancy and boundary artifacts, and offers higher computational efficiency. Therefore, DCT demonstrates significant advantages in infrared and visible image fusion tasks, especially in scenarios that require a balance between global information and local details. However, DCT has certain limitations in capturing non-stationary features and salient target information. To address these issues, we propose a method that combines DCT with attention mechanisms, leveraging DCT's strengths in preserving visible image textures and global information while using attention mechanisms to enhance the capture of salient target information in infrared images, thereby achieving more efficient feature fusion and comprehensive information representation.

3. Methodology

We will give a thorough overview of our visible and infrared image fusion network in this section. First, we will briefly introduce the principles of discrete cosine transform (DCT). Then, we will give an overview of the proposed MSFAFusion network architecture. Next, we will describe in detail the workflow of the MDSCBlock and the SFAFM we designed. Lastly, we shall explain our network's loss function.

3.1. Discrete cosine transform

The discrete cosine transform (DCT) is a commonly used tool in signal processing. DCT was initially developed to overcome some of the limitations of the Fourier transform in handling boundary conditions and real-valued signals. These characteristics make DCT particularly suitable for infrared and visible image fusion, as it can effectively reduce boundary artifacts, concentrate energy in low-frequency components, and preserve critical texture and structural information from both modalities.

The basis functions of the two-dimensional 2D DCT are generally defined as follows:

$$B_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right). \tag{1}$$

The 2D DCT may therefore be expressed as follows:

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{(i,j)}, \tag{2}$$

$$h \in \{0,1,\ldots,H-1\}, \quad w \in \{0,1,\ldots,W-1\}.$$

The 2D DCT spectrum is represented by $f^{2d} \in \mathbb{R}^{H\times W}$, the input is $x^{2d} \in \mathbb{R}^{H\times W}$, the height x^{2d} is represented by H, and the width x^{2d} is represented by W.

3.2. Table of notations

To improve readability, a brief table of notations is provided to clarify the mathematical symbols used in this paper (see Table 2).

Table 2
List of notations.

Symbol	Description
I_{ir}	Input infrared image
I_{vis}	Input visible image
I_{fusion}	Output fused image
I'_{ir}	First input of MDSCBlock,output of Convlayer
φ_{ir}^{i}	Second-fourth input of MDSCBlock
u_i, v_i	The 2D indices of the frequency components corresponding to φ_{ir}^{i}
Ψ^i	Output of the first 3 × 3Conv
$\oplus$	Element-wise addition
I^1, I^2	Residuals of infrared and visible images in CASF
H, W	Image height and width

3.3. Overall network architecture

The primary task of infrared and visible image fusion is to generate a fused image that not only retains the detailed texture information from the visible image but also highlights the salient target information from the infrared image. However, simultaneously preserving the key information from both modalities while achieving satisfactory visual quality remains a significant challenge. In particular, most existing methods are still confined to spatial-domain image fusion, often neglecting the critical information contained in the frequency domain. Therefore, we propose a fusion network called MSFAFusion, which is capable of preserving texture details, highlighting salient target information, and ensuring good visual quality. As illustrated in Fig. 2, the proposed network adopts a streamlined high-level architecture: features are extracted separately from the infrared image I_{ir} and the visible image I_{vi} and the extracted features are subsequently fused in both spatial and frequency domains.

Our network architecture adopts a three-branch design to maximize the retention of key information. The detailed overall framework of the MSFAFusion network is shown in Fig. 3. First, an initial processing module, ConvLayer (Chen et al., 2023), with reflective padding is used to adjust the channel number of the input while preventing boundary information loss, resulting in images I'_{ir} and I'_{vi} that preserve more edge information. Next, this information is fed into the MDSCBlock, which aims to extract multi-scale shallow features from both modalities. The prominent target information characteristics in infrared pictures and the texture information features in visible light images are both efficiently captured by the MDSCBlock. The following is how the features φ_{ir}^{i} and φ_{vi}^{i} that were acquired from the MDSCBlock are processed:

$$\varphi_{ir}^{i} = \begin{cases} MDSCBlock(I'_{ir}), & i=1, \\ MDSCBlock(\varphi_{ir}^{i-1}), & i=2,3,4, \end{cases} \tag{3}$$

the formula that φ_{vi}^{i} follows the same principle as the formula above.

Subsequently, the information processed by the multi-scale depthwise separable convolution block is transmitted to the next module on one hand and to the SFAFM on the other. The SFAFM aims to perform deeper extraction and fusion of the features from infrared and visible images, thereby more effectively highlighting the characteristics of both modalities while retaining more useful information. The process is as follows:

$$\begin{cases} I_{\text{SFAFM}}^{i} = \text{SFAFM}(\varphi_{\text{ir}}^{i}, \varphi_{\text{vi}}^{i}),\ i \in \{1,2,3,4\} \\ I_{\text{SFAFM}} = \text{Concat}(I_{\text{SFAFM}}^{1}, I_{\text{SFAFM}}^{2}, I_{\text{SFAFM}}^{3}, I_{\text{SFAFM}}^{4}) \\ I_{\text{fusion}} = \text{IFP}(\text{DIRM}(\text{Concat}(I_{\text{SFAFM}} \odot \varphi_{\text{ir}}^{i}, I_{\text{SFAFM}} \odot \varphi_{\text{vi}}^{i}))), \end{cases} \tag{4}$$

Finally, the information obtained from the multi-scale depthwise separable convolution block is concatenated with the information processed by the spatial-frequency attention fusion module. The concatenated features are then fed into the dense information reconstruction module (DIRM) and the image fusion path (IFP) (Tang et al., 2023c) to generate the final fused image, which contains enriched information. The architectural details of DIRM and IFP are illustrated in Fig. 4.

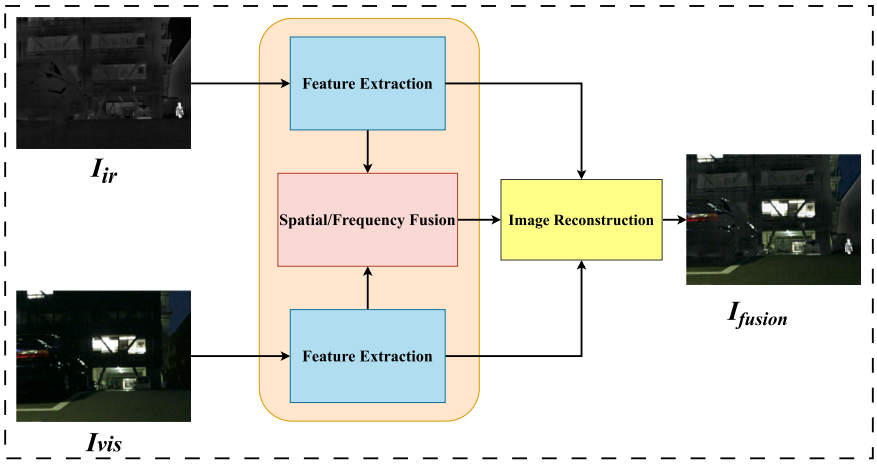


Fig. 2. High-level diagram of the MSFAFusion.

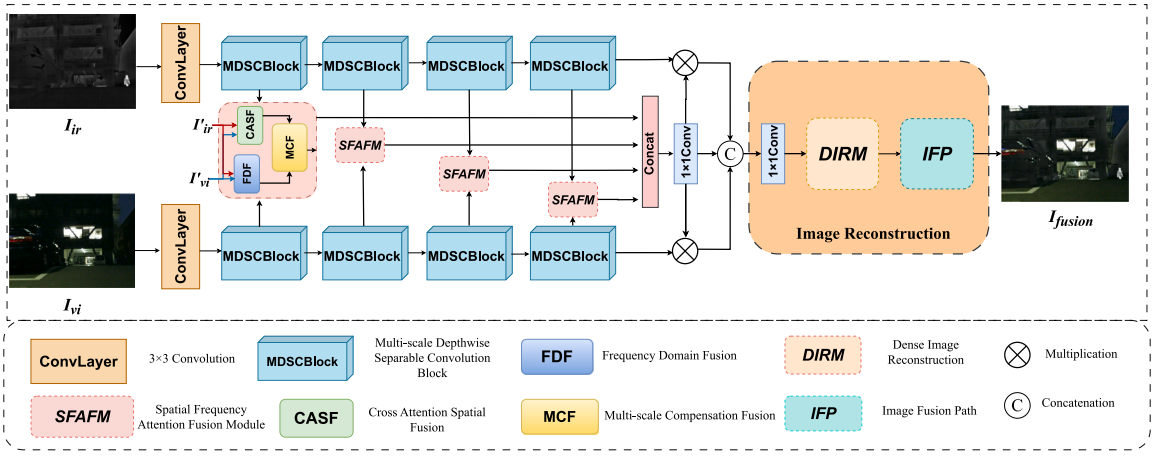


Fig. 3. The proposed MSFAFusion detailed overall architecture consists of the MDSCBlock and the SFAFM. The SFAFM is composed of CASF, FDF, and MCF.

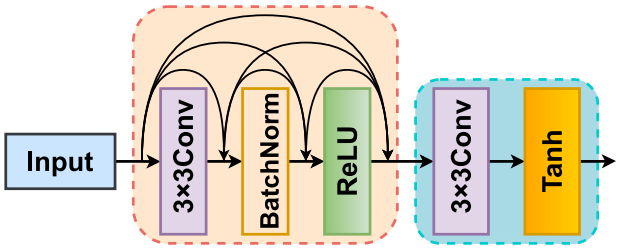


Fig. 4. The architecture of the dense image reconstruction module (DIRM) and the image fusion path (IFP).

Our network can produce fused images directly using the trained fusion network during the testing phase, without the need for any manually created methods.

3.4. Multi-scale depthwise separable convolution block

By combining multi-scale depthwise separable convolution, coordinate attention techniques, pixel normalization, and residual connections, the MDSCBlock is a multi-branch architecture that successfully tackles the feature extraction problem in infrared and visible picture fusion. As shown in Fig. 5, it illustrates the network structure of the MDSCBlock. I'_{ir} and I'_{vi} are the input features that are left over following ConvLayer processing of the source images, which preserves all edge information. First, these two input features are preserved as residual features, denoted as R_{ir}, R_{vi}. The residual connection preserves the original input feature information, which can improve feature learning capability and retain critical information. Each of the other branches uses depthwise separable convolution kernels of different scales (1×1, 3×3, 5×5) for feature extraction, followed by pointwise convolution for dimensionality reduction and feature fusion. By effectively extracting global and local features at various scales, multi-scale depthwise separable convolution (MDSC) improves the model's feature representation capabilities while lowering computational complexity and parameter count, which boosts network efficiency. Based on prior experience, we have designed the number of MDSC modules to be four, which ensures network efficiency while enhancing feature extraction capabilities. The implementation is as follows:

$$I^{i}_{MDSC} = \begin{cases} PConv\left(\text{Concat}(DConv_{1\times1}(I'_{ir}), DConv_{3\times3}(I'_{ir}), DConv_{5\times5}(I'_{ir}))\right) & i = 1 \\ PConv\left(\text{Concat}(DConv_{1\times1}(\varphi^{i-1}_{ir}), DConv_{3\times3}(\varphi^{i-1}_{ir}), DConv_{5\times5}(\varphi^{i-1}_{ir}))\right) & i \in \{2,3,4\} \end{cases} \tag{5}$$

The multi-scale shallow features are concatenated along the channel dimension, and the concatenated features are further processed through the coordinate attention mechanism (CoordAttention) to enhance critical features and integrate cross-channel information. To further harmonize the statistical properties of features from different scales and improve the stability of subsequent feature fusion, pixel normalization (Karras et al., 2018) is applied after multi-scale processing.

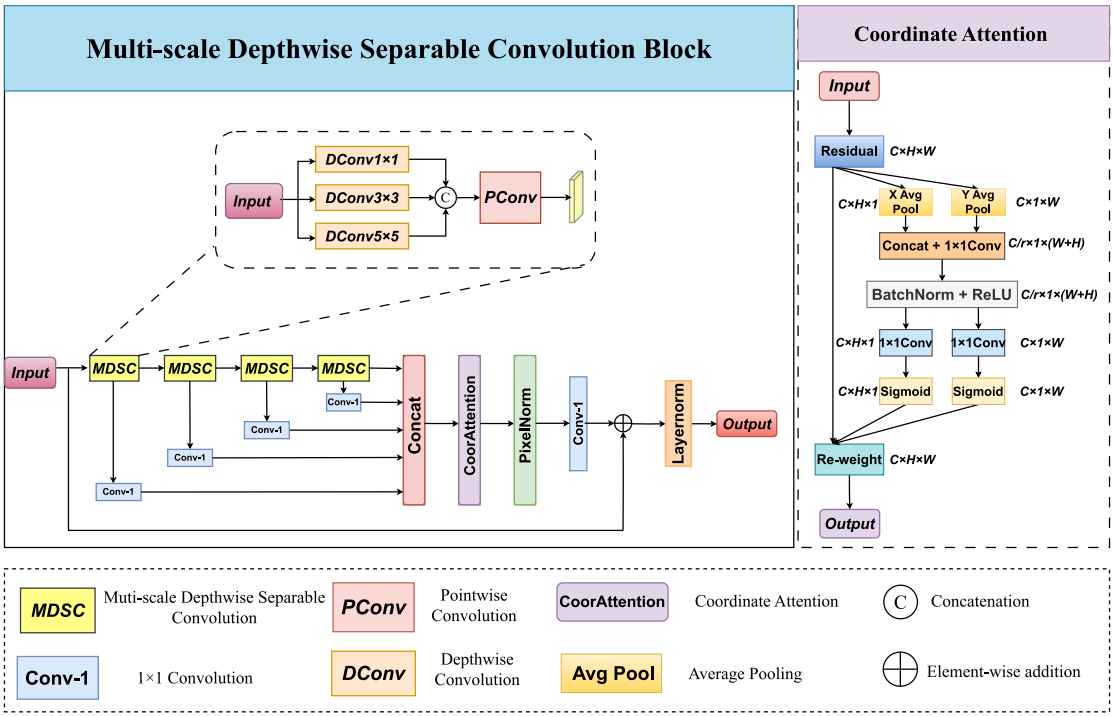


Fig. 5. The detailed architecture of the MDSCBlock we designed.

Finally, the fused features are passed through the LeakyReLU activation function to enhance nonlinear representation capabilities. The process is as follows:

$$I_{MDSCBlock} = \text{LeakyReLU}\left(\text{PixelNorm}\left(\text{CoorAttention} \times \left(\text{Concat}\left(I^{i}_{MDSC}\right)\right)\right) + R_{\text{ir}}\right), i \in \{1,2,3,4\}. \tag{6}$$

The MDSCBlock effectively integrates texture information from visible images and salient target information from infrared images by using convolution kernels of different scales and coordinate attention, enabling better extraction of shallow features from different modality images.

3.5. Spatial-frequency attention fusion module

The SFAFM network structure is divided into three parts: CASF, FDF, and MCF. As shown in Fig. 6(a), (b), and (c), the detailed information of these three fusion modules is presented. As shown in Fig. 6(a),CASF effectively performs spatial-domain interaction between infrared and visible images using cross-attention. It dynamically calculates fusion weights for different modality features. This design fully leverages the complementary information of the two modalities and improves the accuracy of feature representation and fusion performance. The workflow of CASF can be expressed as follows:

$$\Psi^{i} = \text{PatchEmbed}(\varphi^{i}_{ir}), \quad \Psi^{i} \in \mathbb{R}^{\frac{h}{p}\times\frac{w}{p}\times(p^2c)}, i = 1,2,3,4, \tag{7}$$

where h, w, and c indicate the height, width, and channel number of the multi-scale shallow features φ^{i}_{ir} and φ^{i}_{vi}, respectively, and p stands for the patch size. Following layer normalization (LN), the features are mapped to query Q^{i}, key K^{i}, and value $V^{i} \in \mathbb{R}^{\frac{h}{p}\times\frac{w}{p}\times d}$ vectors using three learnable weight matrices W^{Q}, W^{K}, and $W^{V} \in \mathbb{R}^{(p^2c)\times d}$. The following are the representations of Q^{i}, K^{i}, and V^{i}:

$$Q^{i} = \text{LN}(\Psi^{i})W^{Q}, \quad K^{i} = \text{LN}(\Psi^{i})W^{K}, \quad V^{i} = \text{LN}(\Psi^{i})W^{V}, \quad i = 1,2, \tag{8}$$

To calculate the similarity values between the query vectors and key vectors of the two modalities, we use s_1 and s_2 as the similarity values:

$$s_1 = \text{Softmax}\left(\frac{Q^1 K^{1T}}{\sqrt{d}}\right), \quad s_2 = \text{Softmax}\left(\frac{Q^2 K^{2T}}{\sqrt{d}}\right), \tag{9}$$

In order to dynamically weight and fuse the features of visible and infrared images, cross-attention fusion is then used by exchanging similarity scores. The details of the fused image are optimized by constraining the consistency of spatial correlation by crossing the self-similarity maps of the two source images. The expression for the fused features I^{fusion} is:

$$I^{\text{fusion}} = s_1 V^2 \oplus s_2 V^1, \tag{10}$$

where element-wise addition is indicated by $\oplus$. The fused features I^{fusion} are then mapped back to the original feature size using a 3×3 convolution layer. In order to give additional information, residual features I^1 and I^2 are also maintained when processing the original features. The final output I_{CASF} is thus expressed as follows:

$$I_{\text{CASF}} = \text{Conv}(I^{\text{fusion}}) \oplus I^1 \oplus I^2. \tag{11}$$

Inspired by FcaNet (Qin et al., 2021), we propose the Frequency Domain Fusion (FDF) module, which leverages the DCTLayer to compress and integrate visible and infrared features while preserving more informative frequency components. Fig. 6(b) illustrates the input multi-scale shallow features φ^{i}_{ir} are divided into multiple segments along the channel dimension. We denote φ^{i}_{ir} as X^{i}. These segments are denoted as, where $X \in \mathbb{R}^{C'\times H\times W}$, $i \in \{0,1,2,\ldots,n-1\}$, $C' = C/n$, and C must be divisible by n. Each segment is then assigned specific 2D DCT frequency components, and the resulting DCT coefficients are employed as compressed representations for channel attention. The above process can be formulated as:

$$Freq^{i} = \text{2D DCT}^{(u_i,v_i)}(X^{i}) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X^{i}_{h,w} \cdot B^{(u_i,v_i)}_{h,w}, X^{i} = \left[X^{0}, X^{1}, \ldots, X^{n-1}\right], \tag{12}$$

where (u_i, v_i) are the 2D indices of the frequency components corresponding to X^{i}, $\text{Freq}^{i} \in \mathbb{R}^{C'}$ is the compressed C'-dimensional vector.

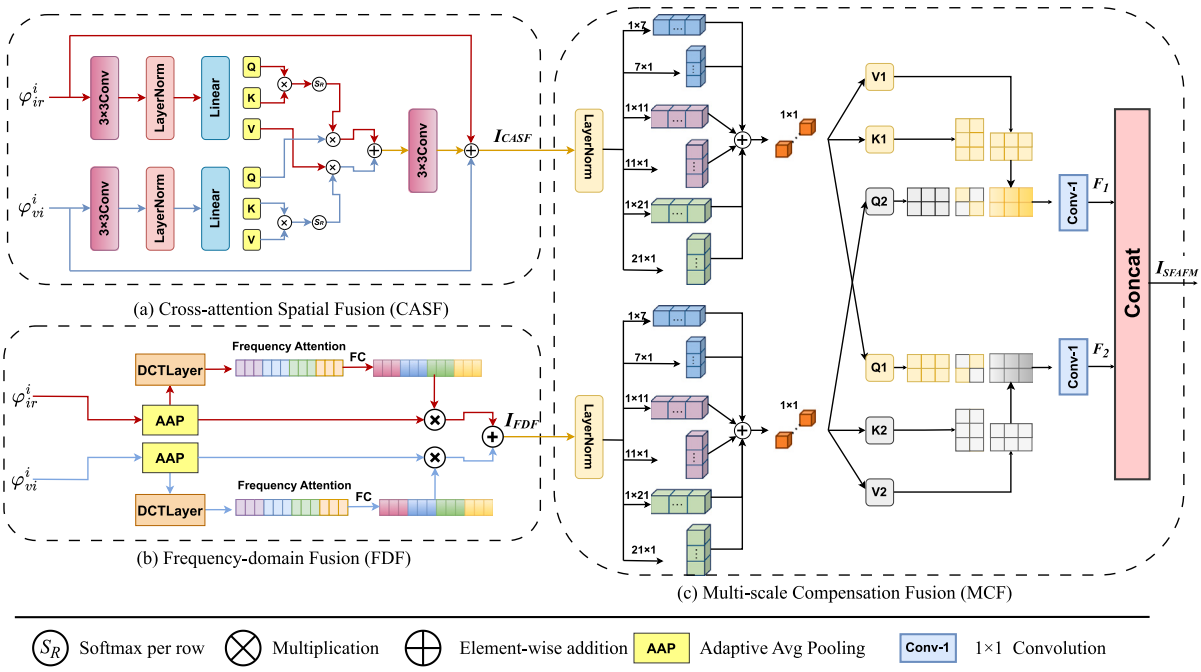


Fig. 6. The detailed information of our designed SFAFM.

The following concatenation procedure can be used to retrieve the full compressed vector:

$$Freq = \text{Concat}([Freq^0, Freq^1, \ldots, Freq^{(n-1)}]), \tag{13}$$

where $Freq^i \in \mathbb{R}^C$ is the vector containing multiple DCT frequency components. A fully connected layer (FC) is then used to further compress the extracted frequency components and enhance their expressive ability. In addition, adaptive average pooling (AAP) is used in the network to compress the feature maps to a resolution suitable for frequency-domain operations in the DCTLayer. At the same time, AAP preserves critical global information during the compression process, providing more effective feature representations for the frequency-domain fusion process. The function of AAP is to adaptively average pool input feature maps of arbitrary size into a specified target size, enabling subsequent operations to align feature dimensions. The entire FDF process can be expressed as:

$$I_{\text{FDF}} = (Freq(\text{AAP}(\varphi_{ir}) \odot \text{AAP}(\varphi_{ir})) \oplus (Freq(\text{AAP}(\varphi_{ir}) \odot \text{AAP}(\varphi_{ir})), \tag{14}$$

Since spatial domain features and frequency domain features capture different attributes and information of the image from distinct dimensions, there exist significant differences at the feature level between the two. These differences can lead to inconsistencies in information representation during feature fusion, thereby affecting the expressive quality of the fusion results. Inspired by SFFNet (Yang et al., 2024), we name the compensation fusion module proposed in this paper as MCF to eliminate the discrepancies between features from different domains. The detailed structure of MCF is shown in Fig. 6(c). First, features from the spatial domain and frequency domain are mapped to a unified scale using multi-scale mapping. Specifically, vertical strip convolutions with different kernel sizes are applied to each feature, followed by concatenation and a 1×1 convolution to map them into unified scale weight matrices Q, K, and V. These processes are applied separately to spatial domain features and frequency domain features, resulting in two sets of weight matrices: $[Q_1, K_1, V_1]$ and $[Q_2, K_2, V_2]$. These two sets of matrices serve as the inputs for the next fusion stage. During the fusion stage, attention mechanisms are employed to compute the key–value pairs of the respective query matrices, followed by feature weighting to achieve feature selection. Finally, the two are concatenated as the final output features. By dynamically adjusting attention weights, the fusion stage ensures that features with similar semantics receive greater attention while features with differing semantics are suppressed, thereby achieving more effective feature compensation. The specific implementation is as follows: For the input spatial domain feature $I_{\text{CASF}} \in \mathbb{R}^{C\times H\times W}$ and frequency domain feature $I_{\text{FDF}} \in \mathbb{R}^{C\times H\times W}$, we define the multi-scale mapping convolutions as:

$$\begin{cases} Q_1, K_1, V_1 = \delta_{1\times1}(OC_7(\text{LN}(I_{\text{CASF}})) + OC_{11}(\text{LN}(I_{\text{CASF}})) \\ \qquad + OC_{21}(\text{LN}(I_{\text{CASF}}))) \\ Q_2, K_2, V_2 = \delta_{1\times1}(OC_7(\text{LN}(I_{\text{FDF}})) + OC_{11}(\text{LN}(I_{\text{FDF}})) \\ \qquad + OC_{21}(\text{LN}(I_{\text{FDF}}))). \end{cases} \tag{15}$$

The attention value is computed as:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{16}$$

where $d = C \times H \times W$. Next, the computation steps in the fusion stage are as follows:

$$\begin{cases} F_1 = \delta_{1\times1}\left(\text{Attn}(Q_2, K_1, V_1)\right) \\ F_2 = \delta_{1\times1}\left(\text{Attn}(Q_1, K_2, V_2)\right). \end{cases} \tag{17}$$

Thus, the output of MCF can be defined as:

$$I_{\text{MCF}} = \text{Concat}(F_1, F_2), \tag{18}$$

where F_1 and F_2 are the feature maps obtained through attention calculations, and $F_1, F_2 \in \mathbb{R}^{\frac{C}{2}\times H\times W}$. F_{MCF} represents the output of the entire MCF, which is also the output of the SFAFM.

3.6. Loss function

The goal of infrared and visible light fusion is to generate a fused image that retains the salient target information from the infrared image while preserving the complex texture details from the visible light image. Therefore, our core objective is to maximize the integration of information between the fused image and the source images. To enhance the overall quality of the fused image, we utilize L_{int} intensity loss and L_{grad} gradient loss to optimize the entire fusion network. The intensity loss primarily measures the differences in intensity values (pixel values) between images, reflecting the pixel-level consistency

between the fused image and the source images. The gradient loss estimates the preservation of structural information in the image by calculating the image gradient, or the rate of change in the image, which assesses the retention of edges or texture characteristics. The definition of intensity loss can be expressed as:

$$L_{int} = \frac{1}{HW}\left\| I_f - \max(I_{ir}, I_{vi}) \right\|_1, \tag{19}$$

Here, H and W represent the height and width of the input image, respectively. $\|\cdot\|_1$ represents l_1-norm, $max(\cdot)$ represents element-wise maximum selection.

The gradient loss is described as follows:

$$L_{\text{grad}} = \frac{1}{HW}\left\| |\nabla I_f| - \max(|\nabla I_{vi}|, |\nabla I_{ir}|) \right\|_1, \tag{20}$$

The fused image is represented by I_f, the L_1-norm used to determine the absolute difference between pixels is indicated by $\|\cdot\|_1$, the maximum selection strategy is indicated by $\max(\cdot)$, the Sobel gradient operator used to measure the image's fine-grained texture is represented by ∇, and the absolute value operation is indicated by $|\cdot|$.

Moreover, we introduced a regularization term L_{corr} to enhance the correlation between the source images and the fused image. The definition of L_{corr} can be expressed as:

$$L_{\text{corr}} = \frac{1}{\text{corr}(I_f, I_{ir}) + \text{corr}(I_f, I_{vis})} \tag{21}$$

Here, $\text{corr}(\cdot)$ represents the calculation of the correlation between two images. Finally, the loss function of MSFAFusion can be expressed as:

$$L_{\text{MSFAFusion}} = \alpha_1 L_{\text{int}} + \alpha_2 L_{\text{grad}} + L_{\text{corr}}, \tag{22}$$

where α_1 and α_2 are hyperparameters that control the balance between different losses.

4. Experiments

The experimental setup and training details are presented first in this section. Then, using generalization and comparing tests, we confirm the efficacy of our approach. In order to confirm the network architecture's rationale, we lastly do ablation investigations, concentrating on the efficiency of the SFAFM and MDSCBlock.

4.1. Experimental configurations

To comprehensively evaluate the proposed MSFAFusion algorithm, we conducted extensive qualitative and quantitative experiments on three widely recognized datasets: MSRS (Tang et al., 2022), Roadscene (Toet and Hogervorst, 2012), and TNO (Xu et al., 2020). These three datasets are among the most commonly used benchmarks in the field of infrared and visible image fusion. They cover diverse scenarios and imaging conditions: MSRS contains a variety of natural scenes with aligned infrared and visible images, RoadScene focuses on urban driving environments under varying lighting conditions, and TNO includes military and surveillance scenes, providing a broad evaluation spectrum. Our suggested approach was contrasted with seven cutting-edge, openly accessible approaches, including one GAN-based approach :GANMcC(2020) (Ma et al., 2020b), two CNN-based methods:U2Fusion(2020) (Xu et al., 2020)and SDNet (2021) (Zhang and Ma, 2021), one image decomposition-based method DeFusion(2022) (Liang et al., 2022), one diffusion model-based method DDFM(2023) (Zhao et al., 2023), one cross-attention-based method CrossFuse (2024) (Li and Wu, 2024), a transformer-based method ITFuse (Tang et al., 2024),one MAE-based method MaeFuse (Li et al., 2025) and one multi-scenario feature joint learning method MLFuse (Lei et al., 2025).

The fusion results were evaluated using seven widely recognized performance evaluation measures. These metrics (Ma et al., 2023) include standard deviation (SD), visual information fidelity (VIF), average gradient (AG), sum of correlation differences (SCD), entropy (EN), gradient-based fusion performance (Qabf), and spatial frequency (SF). Among these, SD gauges the merged image's brightness distribution and contrast; larger values denote better contrast. Better perceptual alignment is indicated by greater VIF values, which quantify the shared information between the fused and source images. AG reflects the richness of texture information, with higher values indicating more detailed textures. SCD measures the information difference between the fused and source images, with higher values reflecting better information retention. EN measures the information content, with higher values indicating richer information. Q_{abf} assesses edge detail preservation, with higher values demonstrating better retention and improved visual quality. SF evaluates spatial frequency richness, with higher values indicating clearer textures and more defined structural features. Furthermore, we measure the segmentation performance using the Intersection over Union (IoU). Better fusion performance is indicated by higher values of these parameters.

4.2. Implementation details

The MSRS dataset, which contains a total of 1444 visible and infrared image pairs with a resolution of 480 × 640 captured in both daytime and nighttime scenarios, was used to train the fusion network. Following a 75%–25% split, 1083 pairs were used for training and 361 pairs for testing. The dataset also provides semantic labels for nine target categories: background, car (or automobile), pedestrian, bicycle, curve, bus stop (or parking station), guardrail (or railing), traffic cone, and speed bump. Prior to being input into the network, all images were normalized to the range [0, 1] and cropped into 64 × 64 patches with a stride of 64 (Tang et al., 2022). This preprocessing yielded a total of 26,112 image patches from both modalities, which were then used to train the fusion model. The images are divided into 64 × 64 patches to enhance data diversity, reduce computational cost, and enable the model to focus more effectively on learning local textures and edge features.

Since the images in our dataset were cropped, the batch size was set to 32, and the number of epochs was set to 400. The model used the Adam optimizer to update the parameters, with an initial learning rate of 0.001 and exponential decay. The hyperparameters for equation (22) were set as $\alpha_1 = 1$, $\alpha_2 = 10$, $\alpha_3 = 1$. This weighting strategy was adopted to emphasize the role of the second term, which corresponds to the detail preservation constraint in the fused image. Empirically, enhancing this term improves the network's ability to retain salient infrared or visible features that are crucial for downstream tasks such as object detection and semantic segmentation. Similar weighting settings have been widely adopted in previous fusion works (Wang et al., 2024), where greater importance is placed on texture or detail-related losses to guide the network in generating more informative fusion results. All experiments were implemented using the PyTorch framework on an NVIDIA Tesla V100 GPU.

4.3. Comparative experiment

To comprehensively evaluate the effectiveness of our method, we compare it with seven other existing state-of-the-art methods on the MSRS dataset.

4.3.1. Qualitative results

We visualized the fused images on the MSRS dataset, as shown in Figs. 7 to 10. These four figures demonstrate that our proposed method, along with nine other algorithms, achieves promising fusion performance. As illustrated in Figs. 7 and 8, in daytime scenes, GANMcC, U2Fusion, and SDNet fail to effectively preserve the texture details of the visible images, resulting in spectral distortion that affects other regions as well. We highlight this issue by enlarging a specific area with a red box. A green box is used to emphasize the degradation of target saliency caused by the introduction of irrelevant information.

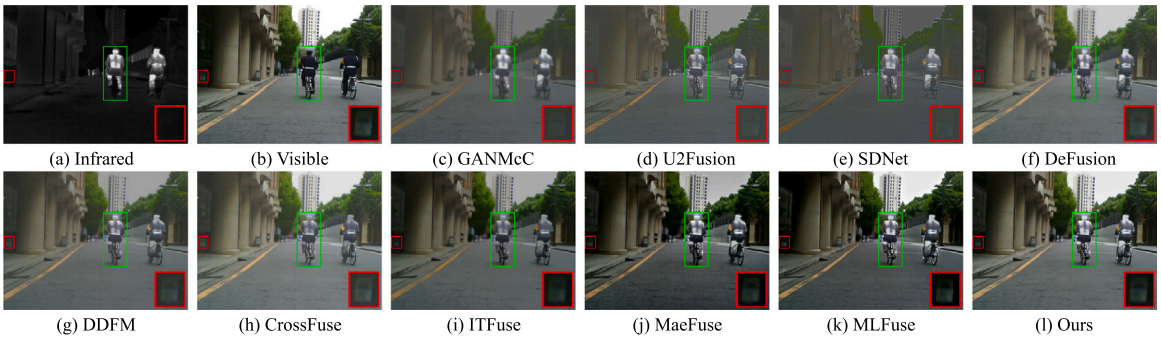


Fig. 7. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the 00537D image from the MSRS dataset.

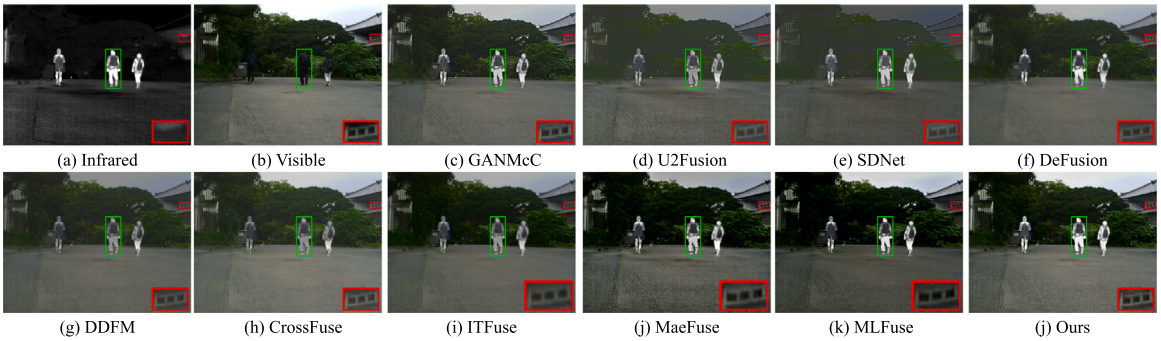


Fig. 8. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the 00634D image from the MSRS dataset.

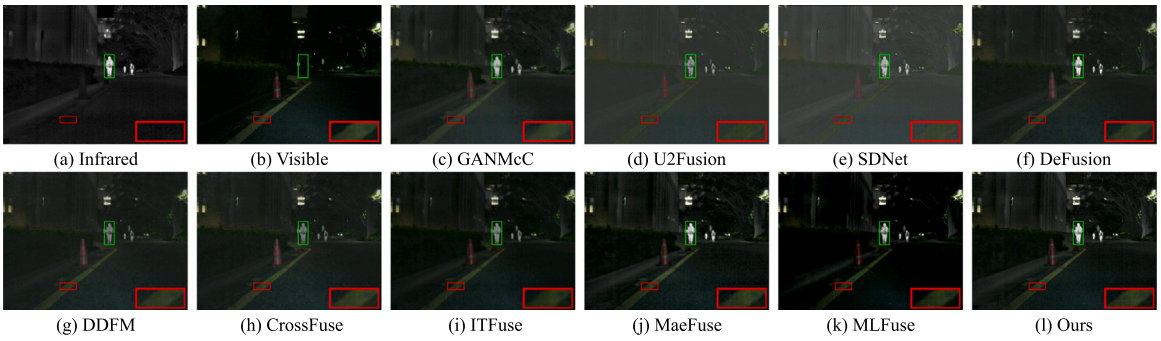


Fig. 9. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the 00858N image from the MSRS dataset.

Only our method, along with MaeFuse and MLFuse, is able to retain rich texture details while enhancing the target information. However, our method demonstrates superior edge sharpness and texture preservation. As shown in Figs. 9 and 10, we use green boxes to mark salient targets. While all methods achieve some degree of fusion between infrared and visible images in nighttime scenes, U2Fusion and SDNet still struggle with spectral distortion and perform poorly under low-light conditions.

4.3.2. Quantitative results

As shown in Table 3, we present the quantitative results on 361 pairs of images evaluated using seven widely adopted metrics. Additionally, we plot the cumulative distribution curves of these metrics to further demonstrate the reliability of our results, as illustrated in Fig. 11. Our method exhibits clear advantages across all seven metrics. SD indicates that our method effectively enhances the contrast and improves the detail representation of the fused images. EN suggests that our method preserves the maximum amount of information from the source images, resulting in the highest information content in the fused images. SCD demonstrates that our method maintains better structural consistency and spatial frequency characteristics between the fused image and the source images. VIF reflects that our fused results are more consistent with the human visual system, offering higher visual fidelity. Qabf indicates that our method performs better in extracting and preserving edge information. AG shows that our approach significantly improves image clarity and edge sharpness. Lastly, SF suggests that our method excels in retaining texture and fine details within the fused image.

4.4. Generalization experiment

We performed generalization experiments on the Roadscene and TNO datasets to confirm our proposed model's capacity for generalization. It is crucial to note that our model was tested directly on the Roadscene and TNO datasets after being trained solely on the MSRS dataset.

4.4.1. Qualitative results

The qualitative results of different compared algorithms on the Roadscene and TNO datasets are shown in Figs. 12 to 15. Figs. 12 and 13 present the qualitative results on the Roadscene dataset. As before,

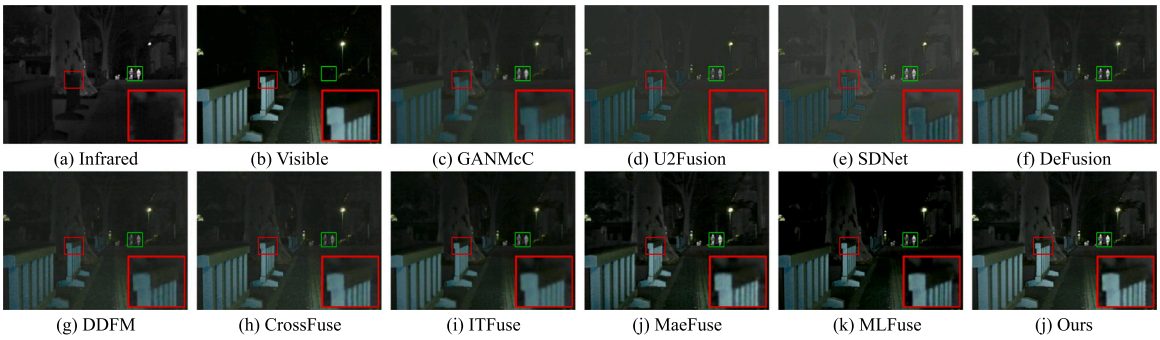


Fig. 10. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the 01024N image from the MSRS dataset.

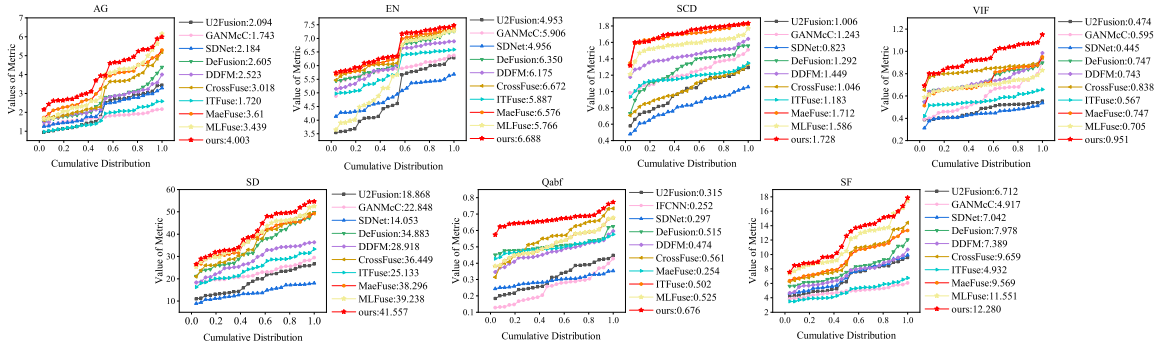


Fig. 11. A quantitative comparison of 361 image pairs from the MSRS dataset across seven metrics (i.e., AG, EN, SCD, VIF, SD, Qabf, and SF).

Table 3
Quantitative comparison of seven metrics (i.e., SD, VIF, AG, SCD, EN, $Qabf$, and SF) on 361 image pairs from the MSRS dataset. Red highlights the top outcomes, and blue highlights the second-best.

	SD	VIF	AG	SCD	EN	Qabf	SF
U2Fusion (Xu et al., 2020)	18.869 ± 7.088	0.474 ± 0.075	2.094 ± 1.040	1.006 ± 0.279	4.953 ± 1.144	0.315 ± 0.100	6.712 ± 2.336
GANMcC (Ma et al., 2020b)	22.848 ± 6.753	0.595 ± 0.163	1.743 ± 0.478	1.243 ± 0.275	5.906 ± 0.450	0.252 ± 0.118	4.917 ± 1.249
SDNet (Zhang and Ma, 2021)	14.053 ± 4.311	0.445 ± 0.068	2.184 ± 0.847	0.823 ± 0.266	4.956 ± 0.619	0.297 ± 0.062	7.042 ± 2.303
DeFusion (Liang et al., 2022)	34.883 ± 11.689	0.747 ± 0.099	2.605 ± 1.041	1.292 ± 0.312	6.350 ± 0.774	0.515 ± 0.064	7.978 ± 2.501
DDFM (Zhao et al., 2023)	28.918 ± 8.661	0.743 ± 0.117	2.523 ± 0.915	1.449 ± 0.219	6.175 ± 0.752	0.474 ± 0.093	7.389 ± 2.231
CrossFuse (Li and Wu, 2024)	36.449 ± 12.045	0.838 ± 0.065	3.018 ± 1.389	1.046 ± 0.293	6.492 ± 0.746	0.561 ± 0.140	9.659 ± 3.358
ITFuse (Tang et al., 2024)	25.133 ± 7.152	0.567 ± 0.057	1.720 ± 0.667	1.183 ± 0.193	5.887 ± 0.764	0.254 ± 0.034	4.932 ± 1.354
MaeFuse (Li et al., 2025)	38.296 ± 10.553	0.747 ± 0.117	3.461 ± 1.317	1.712 ± 0.144	6.576 ± 0.706	0.502 ± 0.050	9.569 ± 2.900
MLFuse (Lei et al., 2025)	39.238 ± 11.601	0.705 ± 0.079	3.439 ± 1.653	1.586 ± 0.186	5.766 ± 1.553	0.525 ± 0.101	11.551 ± 3.878
Ours	41.557 ± 12.339	0.951 ± 0.129	4.003 ± 1.511	1.728 ± 0.150	6.688 ± 0.740	0.676 ± 0.060	12.280 ± 3.838

red boxes are used to highlight texture details, while green boxes indicate salient targets. From the visual analysis, almost all methods are affected by thermal radiation, which leads to the attenuation of salient targets. GANMcC, U2Fusion, SDNet, DeFusion, DDFM, CrossFuse, and MLFuse exhibit particularly noticeable effects of thermal interference. In contrast, our method and MaeFuse are less affected by such issues. The fusion results produced by our method show better consistency with the visible image in background regions, while maintaining pixel intensities for salient targets that closely align with those in the infrared image.

As shown in Figs. 14 and 15 present the qualitative results on the TNO dataset. As in previous figures, red boxes are used to highlight texture details, and green boxes are used to indicate salient targets. U2Fusion, DDFM, and CrossFuse significantly weaken the salient target information during the fusion process, while GANMcC, DeFusion, and DDFM tend to blur fine textures, resulting in reduced image clarity. In contrast, only our method and MaeFuse are capable of simultaneously preserving the texture details from the visible image and the intensity of salient targets. However, MaeFuse retains too many features from the infrared image, which leads to an overemphasis on thermal regions while suppressing some of the environmental texture information from the visible image, ultimately affecting the overall visual balance. In comparison, our method achieves a better trade-off between enhancing salient targets and preserving fine texture details.

4.4.2. Quantitative results

We randomly selected 50 pairs and 25 pairs of images from the Roadscene and TNO datasets, respectively, for quantitative evaluation, and plotted two cumulative distribution curves, as shown in Figs. 16 and 17. The comparative results of the seven evaluation metrics for different algorithms are presented in Tables 4 and 5. As shown in Fig. 16, on the Roadscene dataset, our method achieves the best performance in the VIF, AG, Qabf, and SF metrics, consistently maintaining a leading position. In addition, our method performs competitively on the EN metric, with only a slight gap compared to the best-performing method. From Fig. 17, it can be observed that on the TNO dataset, our method achieves the best results in the SD, AG, EN, Qabf, and SF metrics. Notably, our AG, Qabf, and SF values are the highest among all methods and remain consistently superior. Furthermore, our method also performs closely to the top method in terms of SD and EN metrics.

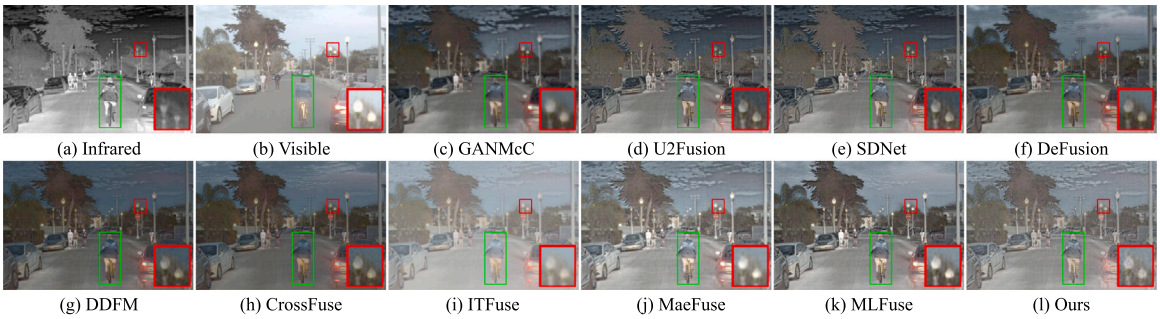


Fig. 12. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the $FLIR\_06832$ image from the RoadScene dataset.

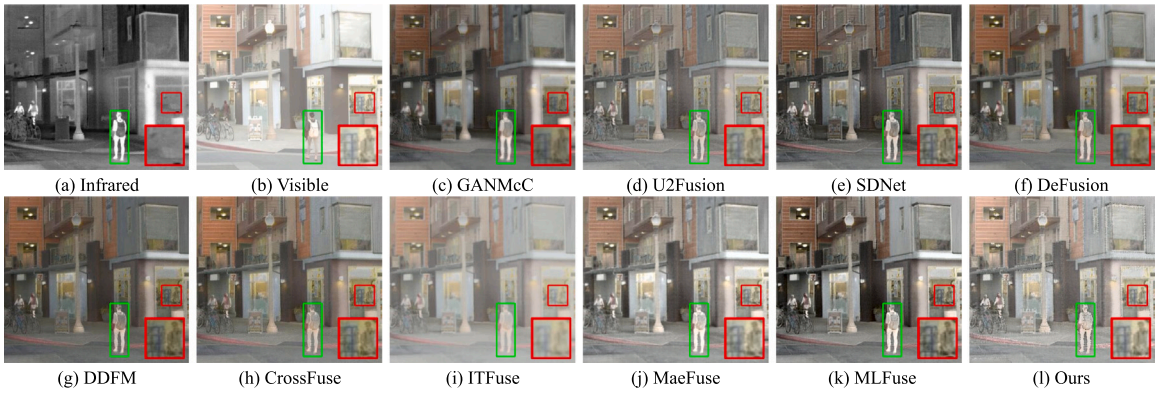


Fig. 13. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the $FLIR\_08835$ image from the Roadscene dataset.

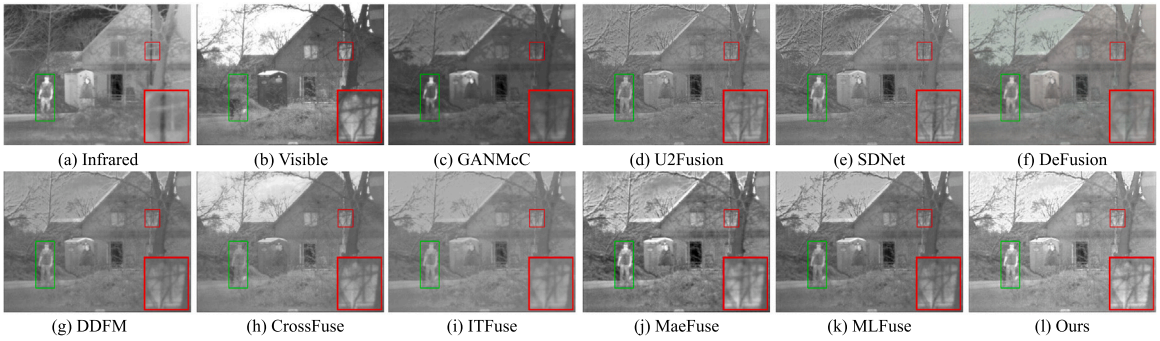


Fig. 14. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the 01 image from the TNO dataset.

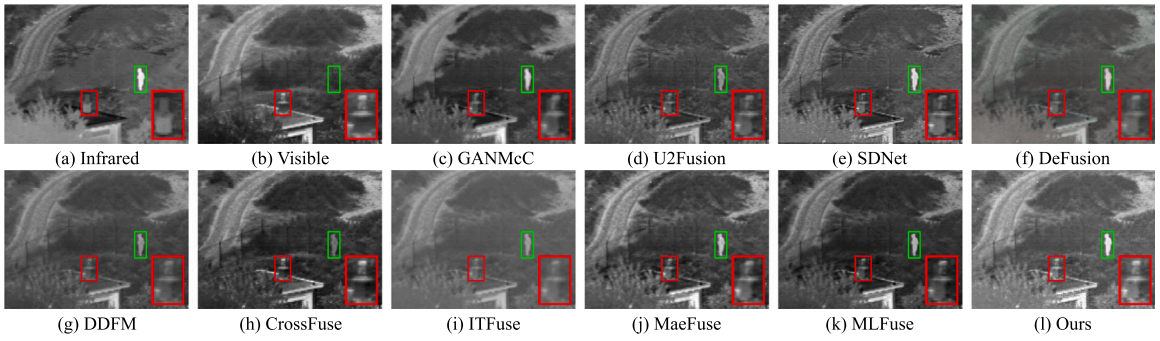


Fig. 15. Qualitative comparison of MSFAFusion with seven state-of-the-art methods on the 19 images from the TNO dataset.

4.5. Task-driven evaluation

More semantic information may be present in the fused images, which enhances performance on challenging vision tasks. In order to assess the segmentation performance of the state-of-the-art techniques we compared, we perform semantic segmentation on the fused images produced by our method. To ensure a fair comparison, we generate fused images from different fusion methods for 1083 pairs of images

Table 4
Quantitative comparison of seven metrics (i.e., SD, VIF, AG, SCD, EN, $Qabf$, and SF) on 50 randomly selected image pairs from the Roadscene dataset. Red highlights the top outcomes, and blue highlights the second-best.

	SD	VIF	AG	SCD	EN	Qabf	SF
U2Fusion (Xu et al., 2020)	29.507 ± 6.792	0.600 ± 0.122	4.109 ± 1.150	1.361 ± 0.444	6.711 ± 0.287	0.491 ± 0.067	10.272 ± 2.884
GANMcC (Ma et al., 2020b)	36.013 ± 7.017	0.469 ± 0.088	3.044 ± 0.659	1.532 ± 0.384	6.926 ± 0.289	0.296 ± 0.071	7.402 ± 1.500
SDNet (Zhang and Ma, 2021)	40.167 ± 8.005	0.600 ± 0.117	5.389 ± 1.518	1.486 ± 0.333	7.140 ± 0.274	0.507 ± 0.068	13.700 ± 3.689
DeFusion (Liang et al., 2022)	34.117 ± 8.295	0.579 ± 0.119	3.063 ± 0.890	1.383 ± 0.378	6.837 ± 0.279	0.414 ± 0.075	7.799 ± 2.256
DDFM (Zhao et al., 2023)	36.161 ± 7.560	0.576 ± 0.121	4.000 ± 1.100	1.704 ± 0.228	7.003 ± 0.260	0.445 ± 0.047	10.555 ± 3.024
CrossFuse (Li and Wu, 2024)	45.371 ± 12.045	0.608 ± 0.114	3.914 ± 1.116	1.306 ± 0.332	7.123 ± 0.229	0.367 ± 0.140	10.746 ± 2.987
ITFuse (Tang et al., 2024)	24.716 ± 5.795	0.427 ± 0.090	1.986 ± 0.497	1.202 ± 0.510	6.326 ± 0.318	0.213 ± 0.036	4.845 ± 1.104
MaeFuse (Li et al., 2025)	39.790 ± 8.762	0.576 ± 0.133	4.508 ± 1.182	1.742 ± 0.171	7.131 ± 0.249	0.460 ± 0.060	10.823 ± 2.750
MLFuse (Lei et al., 2025)	44.330 ± 8.132	0.610 ± 0.126	4.684 ± 1.092	1.837 ± 0.150	7.217 ± 0.226	0.474 ± 0.050	12.976 ± 3.225
Ours	41.201 ± 9.582	0.625 ± 0.163	5.744 ± 1.748	1.665 ± 0.231	7.130 ± 0.241	0.529 ± 0.080	17.236 ± 5.647

Table 5
Quantitative comparison of seven metrics (i.e., SD, VIF, AG, SCD, EN, Qabf, and SF) on 25 randomly selected image pairs from the TNO dataset. Red highlights the top outcomes, and blue highlights the second-best.

	SD	VIF	AG	SCD	EN	Qabf	SF
U2Fusion (Xu et al., 2020)	26.337 ± 9.216	0.539 ± 0.106	3.614 ± 1.198	1.511 ± 0.323	6.417 ± 0.456	0.425 ± 0.053	8.851 ± 3.119
GANMcC (Ma et al., 2020b)	32.187 ± 9.666	0.513 ± 0.114	2.523 ± 0.862	1.634 ± 0.384	6.668 ± 0.399	0.275 ± 0.088	6.217 ± 2.103
SDNet (Zhang and Ma, 2021)	32.650 ± 12.185	0.557 ± 0.132	4.657 ± 1.199	1.492 ± 0.356	6.640 ± 0.361	0.437 ± 0.071	12.056 ± 3.206
DeFusion (Liang et al., 2022)	30.989 ± 12.117	0.553 ± 0.144	2.686 ± 0.987	1.507 ± 0.267	6.582 ± 0.503	0.359 ± 0.101	6.597 ± 2.649
DDFM (Zhao et al., 2023)	32.329 ± 9.031	0.611 ± 0.133	3.238 ± 1.096	1.704 ± 0.280	6.759 ± 0.346	0.427 ± 0.081	8.314 ± 2.661
CrossFuse (Li and Wu, 2024)	40.347 ± 11.624	0.730 ± 0.119	3.777 ± 1.309	1.257 ± 0.391	6.896 ± 0.353	0.459 ± 0.124	10.219 ± 3.024
ITFuse (Tang et al., 2024)	23.247 ± 7.744	0.454 ± 0.111	1.840 ± 0.750	1.403 ± 0.329	6.175 ± 0.451	0.212 ± 0.065	4.279 ± 1.706
MaeFuse (Li et al., 2025)	37.362 ± 11.585	0.544 ± 0.130	3.936 ± 1.565	1.799 ± 0.144	6.994 ± 0.416	0.389 ± 0.063	8.795 ± 3.622
MLFuse (Lei et al., 2025)	36.017 ± 9.887	0.628 ± 0.134	3.878 ± 1.243	1.728 ± 0.237	6.842 ± 0.359	0.468 ± 0.075	10.140 ± 3.091
Ours	38.968 ± 11.660	0.621 ± 0.163	5.170 ± 1.588	1.725 ± 0.161	6.985 ± 0.396	0.532 ± 0.077	15.062 ± 4.038

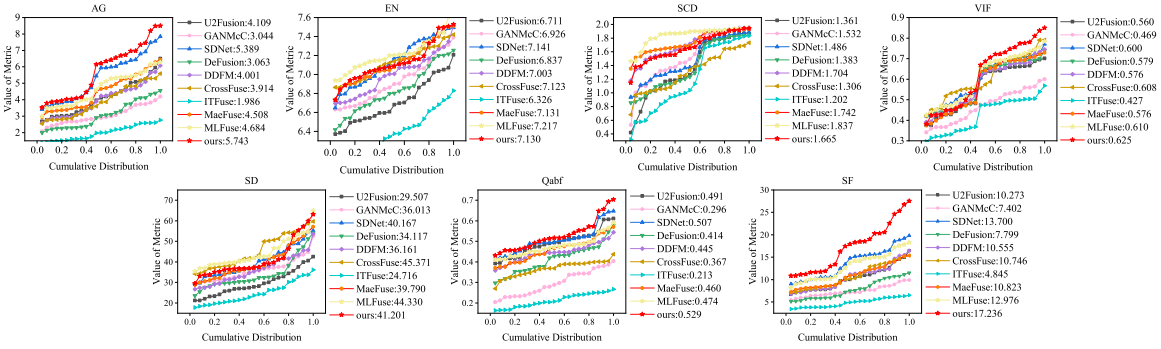


Fig. 16. A quantitative comparison of 361 image pairs from Roadscene dataset across seven metrics (i.e., AG, EN, SCD, VIF, SD, Qabf, and SF).

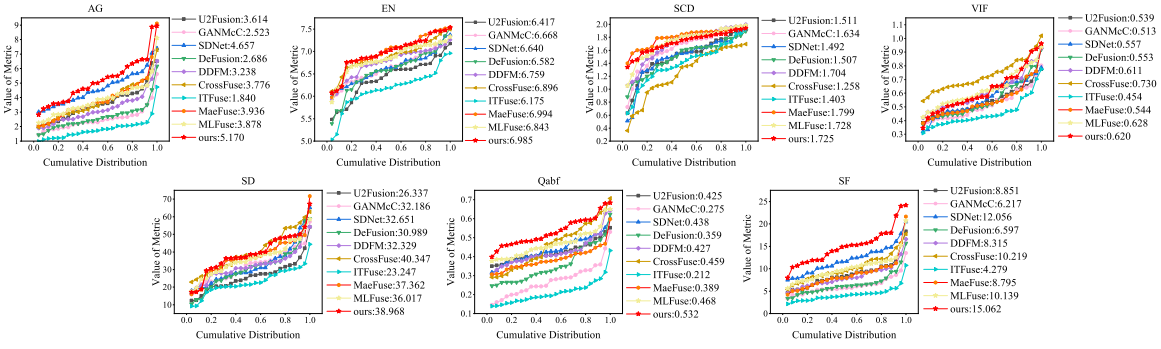


Fig. 17. A quantitative comparison of 361 image pairs from the TNO dataset across seven metrics (i.e., AG, EN, SCD, VIF, SD, Qabf, and SF).

in the training set and 361 pairs in the test set from the MSRS dataset. These two datasets are used for training and testing, respectively. We then retrain the Deeplabv3+ network using the fused images generated from the seven methods on the MSRS training set. The intersection over union (IoU) statistic is used to gauge the segmentation performance. In order to supervise the model training process, we employ cross-entropy and dice loss, and we deploy MobileNetv2 (Sandler et al., 2018) as the backbone network. Stochastic gradient descent with a batch size of four and one hundred epochs is used for training. 7e–3 is the initial learning rate. After training, the fused images from different fusion methods in the test set are passed into the segmentation network for testing, and the segmentation results are shown in Table 6. The segmentation results are also visualized. We can see that our algorithm generally leads in all IoU categories and ranks first in mIoU. This is due to

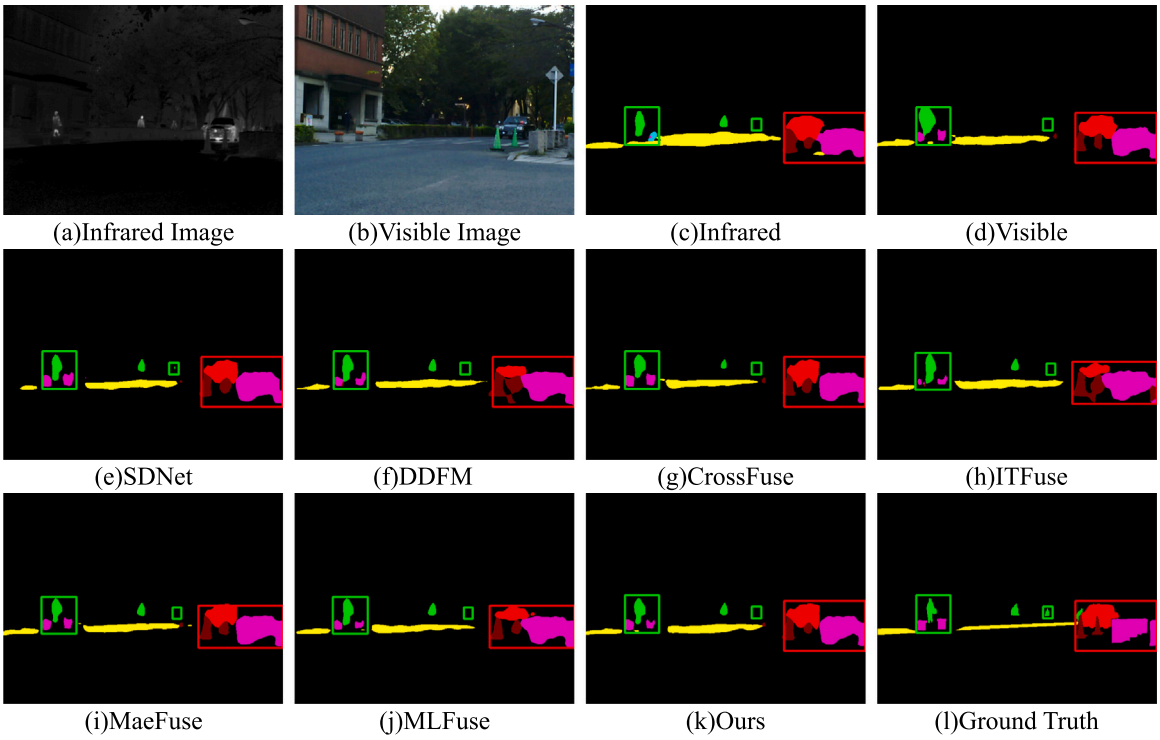


Fig. 18. The segmentation results of the infrared, visible, and fused images for 00127D in the MSRS dataset.

Table 6
The performance of the fusion images from the MSRS dataset in the segmentation model (mIoU). Red highlights the top outcomes, blue highlights the second-best.

	BG	car	Per	Bik	Cur	CS	Gr	CC	Bu	mIoU
Infrared	97.58	85.43	70.18	64.61	50.14	53.02	43.39	45.13	58.02	63.05
Visible	97.85	87.18	59.63	68.54	51.72	66.14	73.27	56.48	65.23	69.56
U2Fusion (Xu et al., 2020)	97.88	87.46	68.79	68.26	51.41	63.67	63.74	53.00	57.88	68.01
GANMcC (Ma et al., 2020b)	97.87	87.09	68.81	67.98	49.44	65.72	64.67	55.40	62.22	68.80
SDNet (Zhang and Ma, 2021)	97.96	87.42	70.45	67.68	53.71	60.50	49.44	54.36	64.43	67.33
DeFusion (Liang et al., 2022)	97.87	87.09	69.06	66.29	51.66	62.83	72.55	50.30	62.50	68.91
DDFM (Zhao et al., 2023)	97.79	86.60	66.01	67.53	49.56	64.14	64.53	46.25	63.52	67.33
CrossFuse (Li and Wu, 2024)	97.90	86.26	64.21	67.21	48.76	65.65	74.17	55.56	64.76	69.39
ITFuse (Tang et al., 2024)	97.69	86.43	63.62	66.00	49.58	61.74	55.85	45.88	61.51	65.37
MaeFuse (Li et al., 2025)	97.76	85.33	69.36	67.66	50.23	59.50	40.04	55.43	62.06	65.27
MLFuse (Lei et al., 2025)	97.53	83.25	66.09	66.77	45.77	62.04	40.06	50.49	54.53	62.95
Ours	98.06	87.84	69.13	68.08	54.45	67.51	75.29	56.34	66.60	71.48

the efficient use of the DCT method's characteristics in our network, demonstrating excellent texture information handling capabilities. Additionally, by using the attention mechanism, we further enhance the network's ability to capture significant target information and interact with features from two different domains. The segmentation network may better represent the imaging scene by strengthening both spatial and semantic information under the guidance of semantic loss.

As shown in Fig. 18, we present the visualized segmentation results. For comparison, we include the segmentation outputs of SDNet, DDFM, CrossFuse, ITFuse, MaeFuse, and MLFuse. The results demonstrate that the segmentation model used in our framework achieves promising performance. For instance, in the image 00127D, our method enables more accurate segmentation of pedestrians and background regions.

4.6. Ablation study

To validate the effectiveness of the proposed MDSCBlock and SFAFM, we conducted ablation experiments with visual analysis, as shown in Fig. 19. As illustrated in Table 7, the removal of either module results in a decline in performance metrics. Specifically, the removal of MDSCBlock or SFAFM weakens the model's ability to extract features and integrate information from the source images. While SD shows only a slight drop, the significant declines in SCD, AG, and EN confirm this degradation. The decrease in SCD indicates a larger discrepancy between the fused image and the source images, implying a loss of structural information. A decline in AG reveals a loss of texture detail, while the reduced EN suggests diminished information content in the fused image, reflecting poorer integration capabilities.

Due to the relatively small difference in SD values, contrast variations in the visual results are not very noticeable. The magnified regions in Fig. 19 reveal clear differences in edge quality. Our method preserves edge information around the streetlamp more effectively. The W/O SFAFM and W/O CASF versions clearly exhibit edge blurring and fragmentation, indicating that SFAFM plays a critical role in enhancing structural perception in salient regions. Additionally, removing the MCF or FDF modules leads to overexposure or detail loss in the highlighted (red box) illumination areas, suggesting these modules significantly contribute to dynamic range adjustment and frequency-domain compensation.

When comparing W/O MDSCBlock(1,3,7) with W/O MDSCBlock(3,5,7), the former shows poor texture fusion across multiple scales, with noticeable edge blurring in the streetlamp area.

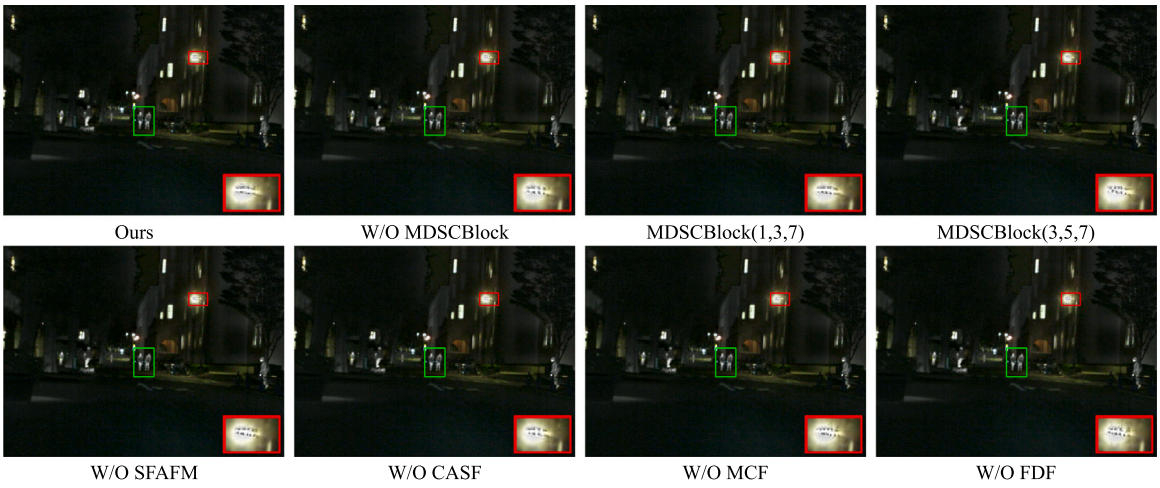


Fig. 19. Visualized results of ablation.

Table 7
The evaluation metrics of ablation study. Red indicates the best result and blue represents the second best result.

	SD	AG	SCD	EN
W/o MDSCBlock	40.346 ± 11.748	3.956 ± 1.498	1.659 ± 0.176	6.635 ± 0.750
MDSCBlock(1,3,7)	40.984 ± 12.154	4.002 ± 1.498	1.728 ± 0.147	6.661 ± 0.747
MDSCBlock(3,5,7)	41.121 ± 12.175	3.995 ± 1.520	1.713 ± 0.155	6.664 ± 0.748
W/o SFAFM	41.540 ± 12.194	3.986 ± 1.527	1.712 ± 0.149	6.672 ± 0.748
W/o CASF	40.369 ± 12.078	3.977 ± 1.505	1.657 ± 0.164	6.593 ± 0.794
W/o FDF	41.344 ± 12.228	3.993 ± 1.514	1.705 ± 0.160	6.657 ± 0.771
W/o MCF	40.996 ± 12.277	3.988 ± 1.510	1.678 ± 0.172	6.625 ± 0.788
Ours	41.557 ± 12.339	4.002 ± 1.511	1.728 ± 0.151	6.688 ± 0.740

Although the latter performs better in terms of edge clarity, it does not adequately preserve the texture details of the lamp, indicating that larger kernel sizes do not necessarily yield better results. This comparison demonstrates that different kernel scales have different strengths in capturing local texture versus global structure. Therefore, we adopt the (1,3,5) configuration in MDSCBlock: the small-scale (1×1) kernel extracts fine-grained texture, the mid-scale (3×3) kernel captures local structural relationships, and the large-scale (5×5) kernel enhances contextual awareness. This combination strikes a balance between image clarity, structural integrity, and information richness, ultimately leading to superior fusion performance. These differences, particularly in the highlighted regions, further verify the effectiveness of our proposed method.

5. Conclusion

This paper proposes a multi-scale frequency attention fusion network called MSFAFusion for infrared and visible image fusion. By designing MDSCBlock and SFAFM, our model effectively extracts and fuses multi-scale feature information from different modality images, preserving the texture details of the visible light images while emphasizing the significant target information from the infrared images. According to experimental results, MSFAFusion performs significantly better than state-of-the-art fusion techniques on a number of publicly available datasets, with notable benefits in both quantitative metrics and qualitative visualizations. Furthermore, generalization experiments and task-driven experiments further validate the robustness of our model and its potential for real-world applications.

Despite its good performance, MSFAFusion still has some limitations. For example, the model has high computational complexity, which limits its efficiency in real-time applications. The network's dependency on specific scenes may affect its performance in more complex or unknown scenarios. In addition, there is still room for improvement in the extraction of significant target information. Future work could focus on optimizing the network structure to reduce computational overhead while exploring more data-driven training strategies to enhance the model's adaptability and generalization capabilities.

CRediT authorship contribution statement

Yong Wang: Supervision, Funding acquisition. **Xueyuan Zhao:** Writing – original draft, Visualization, Methodology. **Jianfei Pu:** Validation, Formal analysis. **Lulu Zhang:** Project administration, Investigation. **Duoqian Miao:** Funding acquisition.

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.

Acknowledgments

The National Natural Science Foundation of China provided funding for this work under Grant 62376198.

Data availability

Data will be made available on request.

References

Akhtar, M.U., Liu, J., Xie, Z., Cui, X., Liu, X., Huang, B., 2025. Multilingual entity alignment by abductive knowledge reasoning on multiple knowledge graphs. Eng. Appl. Artif. Intell. 139, 109660. http://dx.doi.org/10.1016/j.engappai.2024.109660.

Chang, Z., Feng, Z., Yang, S., Gao, Q., 2023. AFT: Adaptive fusion transformer for visible and infrared images. IEEE Trans. Image Process. 32, 2077–2092. http://dx.doi.org/10.1109/TIP.2023.3263113.

Chen, J., Ding, J., Yu, Y., Gong, W., 2023. THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor. Neurocomputing 527, 71–82. http://dx.doi.org/10.1016/j.neucom.2023.01.033.

Cvejic, N., Bull, D., Canagarajah, N., 2007. Region-based multimodal image fusion using ICA bases. IEEE Sens. J. 7 (5), 743–751. http://dx.doi.org/10.1109/icif.2006.301600.

Ehrlich, M., Davis, L.S., 2019. Deep residual learning in the jpeg transform domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3484–3493. http://dx.doi.org/10.48550/arXiv.1812.11690.

Gao, L., Zhuang, Z., Tao, H., Chen, Y., Stojanovic, V., 2024. Non-lifted norm optimal iterative learning control for networked dynamical systems: A computationally efficient approach. J. Franklin Inst. 361 (15), 107112. http://dx.doi.org/10.1016/j.jfranklin.2024.107112.

Gueguen, L., Sergeev, A., Kadlec, B., Liu, R., Yosinski, J., 2018. Faster neural networks straight from JPEG. In: Advances in Neural Information Processing Systems. Vol. 31, Curran Associates, Inc..

Han, D., Li, L., Guo, X., Ma, J., 2022. Multi-exposure image fusion via deep perceptual enhancement. Inf. Fusion 79, 248–262. http://dx.doi.org/10.1016/j.inffus.2021.10.006.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722. http://dx.doi.org/10.48550/arXiv.2103.02907.

Jain, D.K., Zhao, X., González-Almagro, G., Gan, C., Kotecha, K., 2023. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. Inf. Fusion 95, 401–414. http://dx.doi.org/10.1016/j.inffus.2023.02.014.

Jerripothula, K.R., Mukherjee, P., Cai, J., Lu, S., Yuan, J., 2022. AppFuse: An appearance fusion framework for saliency cues. IEEE Trans. Circuits Syst. Video Technol. 32 (12), 8261–8274. http://dx.doi.org/10.1109/TCSVT.2022.3188699.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. http://dx.doi.org/10.48550/arXiv.1710.10196, arXiv:1710.10196 [cs.NE].

Khayam, S.A., 2003. The discrete cosine transform (DCT): theory and application. Mich. State Univ. 114 (1), 31.

Lei, J., Li, J., Liu, J., Wang, B., Zhou, S., Zhang, Q., Wei, X., Kasabov, N.K., 2025. MLFuse: Multi-scenario feature joint learning for Multi-Modality image fusion. IEEE Trans. Multimed. 1–16. http://dx.doi.org/10.1109/TMM.2025.3535355.

Li, J., Jiang, J., Liang, P., Ma, J., Nie, L., 2025. MaeFuse: Transferring Omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training. IEEE Trans. Image Process. 34, 1340–1353. http://dx.doi.org/10.1109/TIP.2025.3541562.

Li, H., Wu, X.-J., 2018. DenseFuse: A fusion approach to infrared and visible images. IEEE Trans. Image Process. 28 (5), 2614–2623. http://dx.doi.org/10.1109/tip.2018.2887342.

Li, H., Wu, X.-J., 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. Inf. Fusion 103, 102147. http://dx.doi.org/10.1016/j.inffus.2023.102147.

Li, H., Wu, X.-J., Durrani, T., 2020a. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans. Instrum. Meas. 69 (12), 9645–9656. http://dx.doi.org/10.1109/tim.2020.3005230.

Li, H., Wu, X.-J., Kittler, J., 2020b. MDLatLRR: A novel decomposition method for infrared and visible image fusion. IEEE Trans. Image Process. 29, 4733–4746. http://dx.doi.org/10.1109/tip.2020.2975984.

Liang, P., Jiang, J., Liu, X., Ma, J., 2022. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In: European Conference on Computer Vision. Springer, pp. 719–735. http://dx.doi.org/10.1007/978-3-031-19797-0_41.

Lin, H., Cheng, X., Wu, X., Shen, D., 2022. Cat: Cross attention in vision transformer. In: 2022 IEEE International Conference on Multimedia and Expo. ICME, pp. 1–6. http://dx.doi.org/10.1109/ICME52920.2022.9859720.

Liu, Y., Chen, X., Ward, R.K., Wang, Z.J., 2016. Image fusion with convolutional sparse representation. IEEE Signal Process. Lett. 23 (12), 1882–1886. http://dx.doi.org/10.1109/lsp.2016.2618776.

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. http://dx.doi.org/10.48550/arXiv.2203.16220.

Liu, Y., Jin, J., Wang, Q., Shen, Y., Dong, X., 2014. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. Signal Process. 97, 9–30. http://dx.doi.org/10.1016/j.sigpro.2013.10.010.

Long, Y., Jia, H., Zhong, Y., Jiang, Y., Jia, Y., 2021. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. Inf. Fusion 69, 128–141. http://dx.doi.org/10.1016/j.inffus.2020.11.009.

Ma, J., Chen, C., Li, C., Huang, J., 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. Inf. Fusion 31, 100–109. http://dx.doi.org/10.1016/j.inffus.2016.02.001.

Ma, J., Liang, P., Yu, W., Chen, C., Guo, X., Wu, J., Jiang, J., 2020a. Infrared and visible image fusion via detail preserving adversarial learning. Inf. Fusion 54, 85–98. http://dx.doi.org/10.1016/j.inffus.2019.07.005.

Ma, J., Ma, Y., Li, C., 2019b. Infrared and visible image fusion methods and applications: A survey. Inf. Fusion 45, 153–178. http://dx.doi.org/10.1016/j.inffus.2018.02.004.

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 J. Autom. Sin. 9 (7), 1200–1217. http://dx.doi.org/10.1109/JAS.2022.105686.

Ma, J., Tang, L., Xu, M., Zhang, H., Xiao, G., 2021. STDFusionNet: An infrared and visible image fusion network based on salient target detection. IEEE Trans. Instrum. Meas. 70, 1–13. http://dx.doi.org/10.1109/tim.2021.3075747.

Ma, W., Wang, K., Li, J., Yang, S.X., Li, J., Song, L., Li, Q., 2023. Infrared and visible image fusion technology and application: A review. Sens. 23 (2), 599. http://dx.doi.org/10.3390/s23020599.

Ma, J., Yu, W., Liang, P., Li, C., Jiang, J., 2019a. FusionGAN: A generative adversarial network for infrared and visible image fusion. Inf. Fusion 48, 11–26. http://dx.doi.org/10.1016/j.inffus.2018.09.004.

Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H., 2020b. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE Trans. Instrum. Meas. 70, 1–14. http://dx.doi.org/10.1109/TIM.2020.3038013.

Maruschak, P., Konovalenko, I., Osadtsa, Y., Medvid, V., Shovkun, O., Baran, D., Kozbur, H., Mykhailyshyn, R., 2024. Surface illumination as a factor influencing the efficacy of defect recognition on a rolled metal surface using a deep neural network. Appl. Sci. 14 (6), http://dx.doi.org/10.3390/app14062591.

Ning, M., Ze-Ming, Z., ZHANG, P., Li-Min, L., 2013. A new variational model for panchromatic and multispectral image fusion. Acta Automat. Sinica 39 (2), 179–187. http://dx.doi.org/10.1016/S1874-1029(13)60020-8.

Paramanandham, N., Rajendiran, K., 2018. Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. Infrared Phys. Technol. 88, 13–22. http://dx.doi.org/10.1016/j.infrared.2017.11.006.

Qin, Z., Zhang, P., Wu, F., Li, X., 2021. Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 783–792. http://dx.doi.org/10.48550/arXiv.2012.11879.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520. http://dx.doi.org/10.48550/arXiv.1801.04381.

Sifuzzaman, M., Islam, M.R., Ali, M.Z., 2009. Application of wavelet transform and its advantages compared to Fourier transform.

Sun, Y., Tao, H., Stojanovic, V., 2025. End-to-end multi-scale residual network with parallel attention mechanism for fault diagnosis under noise and small samples. ISA Trans. 157, 419–433. http://dx.doi.org/10.1016/j.isatra.2024.12.023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Tang, W., He, F., Liu, Y., 2023a. TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation. Pattern Recognit. 137, 109295. http://dx.doi.org/10.1016/j.patcog.2022.109295.

Tang, W., He, F., Liu, Y., 2024. ITFuse: An interactive transformer for infrared and visible image fusion. Pattern Recognit. 156, 110822. http://dx.doi.org/10.1016/j.patcog.2024.110822.

Tang, L., Xiang, X., Zhang, H., Gong, M., Ma, J., 2023b. DIVFusion: Darkness-free infrared and visible image fusion. Inf. Fusion 91, 477–493. http://dx.doi.org/10.1016/j.inffus.2022.10.034.

Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J., 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. Inf. Fusion 83, 79–92. http://dx.doi.org/10.1016/j.inffus.2022.03.007.

Tang, L., Zhang, H., Xu, H., Ma, J., 2023c. 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. Inf. Fusion 99, 101870. http://dx.doi.org/10.1016/j.inffus.2023.101870.

Tao, H., Zheng, Y., Wang, Y., Qiu, J., Stojanovic, V., 2024. Enhanced feature extraction YOLO industrial small object detection algorithm based on receptive-field attention and multi-scale features. Meas. Sci. Technol. 35 (10), 105023. http://dx.doi.org/10.1088/1361-6501/ad633d.

Toet, A., 1989. Image fusion by a ratio of low-pass pyramid. Pattern Recognit. Lett. 9 (4), 245–253. http://dx.doi.org/10.1016/0167-8655(89)90003-2.

Toet, A., Hogervorst, M.A., 2012. Progress in color night vision. Opt. Eng., Bellingham 51 (1), http://dx.doi.org/10.1117/1.OE.51.1.010901, 010901–010901.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. 30, Curran Associates, Inc., http://dx.doi.org/10.48550/arXiv.1706.03762.

Wang, Y., Pu, J., Miao, D., Zhang, L., Zhang, L., Du, X., 2024. SCGRFuse: An infrared and visible image fusion network based on spatial/channel attention mechanism and gradient aggregation residual dense blocks. Eng. Appl. Artif. Intell. 132, 107898. http://dx.doi.org/10.1016/j.engappai.2024.107898.

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H., 2020. U2Fusion: A unified unsupervised image fusion network. IEEE Trans. Pattern Anal. Mach. Intell. 44 (1), 502–518. http://dx.doi.org/10.1109/tpami.2020.3012548.

Yang, Y., Yuan, G., Li, J., 2024. SFFNet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. IEEE Trans. Geosci. Remote Sens. 62, 1–17. http://dx.doi.org/10.1109/TGRS.2024.3427370.

Yu, H., Zheng, N., Zhou, M., Huang, J., Xiao, Z., Zhao, F., 2022. Frequency and spatial dual guidance for image dehazing. In: European Conference on Computer Vision. pp. 181–198. http://dx.doi.org/10.1007/978-3-031-19800-7_11.

Zhang, B., 2010. Study on image fusion based on different fusion rules of wavelet transform. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering. ICACTE, Vol. 3, pp. V3–649. http://dx.doi.org/10.1109/ICACTE.2010.5579586.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L., 2020a. IFCNN: A general image fusion framework based on convolutional neural network. Inf. Fusion 54, 99–118. http://dx.doi.org/10.1016/j.inffus.2019.07.011.

Zhang, H., Ma, J., 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. Int. J. Comput. Vis. 129 (10), 2761–2785. http://dx.doi.org/10.1007/s11263-021-01501-8.

Zhang, H., Ma, X., Tian, Y., 2020b. An image fusion method based on curvelet transform and guided filter enhancement. Math. Probl. Eng. 2020 (1), 9821715. http://dx.doi.org/10.1155/2020/9821715.

Zhang, P., Wang, D., Lu, H., Yang, X., 2021a. Learning adaptive attribute-driven representation for real-time RGB-T tracking. Int. J. Comput. Vis. 129, 2714–2729. http://dx.doi.org/10.1007/s11263-021-01495-3.

Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., Han, J., 2021b. ABMDRNet: Adaptive-weighted Bi-Directional modality difference reduction network for RGB-T semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2633–2642.

Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L., 2023. DDFM: denoising diffusion model for multi-modality image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8082–8093. http://dx.doi.org/10.48550/arXiv.2303.06840.

Zheng, N., Zhou, M., Huang, J., Zhao, F., 2024. Frequency integration and spatial compensation network for infrared and visible image fusion. Inf. Fusion 109, 102359. http://dx.doi.org/10.1016/j.inffus.2024.102359.

Zhou, M., Huang, J., Yan, K., Hong, D., Jia, X., Chanussot, J., Li, C., 2025. A general spatial-frequency learning framework for multimodal image fusion. IEEE Trans. Pattern Anal. Mach. Intell. 47 (7), 5281–5298. http://dx.doi.org/10.1109/TPAMI.2024.3368112.

Zhou, X., Jiang, Z., Okuwobi, I.P., 2023. CAFNET: Cross-attention fusion network for infrared and low illumination visible-light image. Neural Process. Lett. 55 (5), 6027–6041. http://dx.doi.org/10.1007/s11063-022-11125-9.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J., 2021a. DeepViT: Towards deeper vision transformer. http://dx.doi.org/10.48550/arXiv.2103.11886, arXiv:2103.11886 [cs.CV].

Zhou, H., Wu, W., Zhang, Y., Ma, J., Ling, H., 2021b. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. IEEE Trans. Multimed. 25, 635–648. http://dx.doi.org/10.1109/TMM.2021.3129609.