ICJK

Chinese Journal of Information Fusion
http://dx.doi.org/10.62762/CJIF.2024.655617

RESEARCH ARTICLE

DMFuse: Diffusion Model Guided Cross-Attention Learning for Infrared and Visible Image Fusion

Wuqiang Qi[1], **Zhuoqun Zhang**[1] **and Zhishe Wang**[1,*]

[1] School of Applied Science, Taiyuan University of Science and Technology, Taiyuan 030024, China

Abstract

Image fusion aims to integrate complementary information from different sensors into a single fused output for superior visual description and scene understanding. The existing GAN-based fusion methods generally suffer from multiple challenges, such as unexplainable mechanism, unstable training, and mode collapse, which may affect the fusion quality. To overcome these limitations, this paper introduces a diffusion model guided cross-attention learning network, termed as DMFuse, for infrared and visible image fusion. Firstly, to improve the diffusion inference efficiency, we compress the quadruple channels of the denoising UNet network to achieve more efficient and robust model for fusion tasks. After that, we employ the pre-trained diffusion model as an autoencoder and incorporate its strong generative priors to further train the following fusion network. This design allows the generated diffusion features to effectively showcase high-quality distribution mapping ability. In addition, we devise a cross-attention interactive fusion module to establish the long-range dependencies from local diffusion features. This module integrates the global interactions to improve the complementary characteristics of different modalities. Finally, we propose a multi-level decoder network to reconstruct the fused output. Extensive experiments on fusion tasks and downstream applications, including object detection and semantic segmentation, indicate that the proposed model yields promising performance while maintaining competitive computational efficiency. The code and data are available at https://github.com/Zhishe-Wang/DMFuse.

Keywords: image fusion, diffusion model, feature interaction, attention mechanism, deep generative model.

1 Introduction

Infrared sensors detect hidden target characteristics through thermal radiation and work under various weather and lighting conditions. The acquired images are often exhibit low contrast and lack fine details. On the contrary, visible sensors offer high-resolution scene perception through light reflection imaging. However, under adverse weather or camouflage conditions, visible sensors are difficult to distinguish obvious targets from the background environment. The image fusion technology can integrate the complementary information from different sensors into a single image,

Academic Editor:
Jun Shen

Submitted: 24 August 2024
Accepted: 28 December 2024
Published: 31 December 2024

Vol. 1, **No.** 3, 2024.
10.62762/CJIF.2024.655617

***Corresponding author:**
Zhishe Wang
wangzs@tyust.edu.cn

Citation
Qi, W., Zhang, Z., & Wang, Z. (2024). DMFuse: Diffusion Model Guided Cross-Attention Learning for Infrared and Visible Image Fusion. *Chinese Journal of Information Fusion*, 1(3), 226–242.

which can achieve superior visual description and scene understanding. A common application of fused images is to provide faster and more accurate visual interpretation for both human observers and computer systems. In addition, this technology has been extended into other visual tasks, such as person re-identification [1], object detection [2], and RGBT tracking [3], and so on.

Over the past decades, traditional algorithms, including multi-scale transformation [4], sparse representation [5], subspace decomposition [6], optimization model [7], hybrid-based [8], and other methods [9], have been proposed for infrared and visible image fusion. Although these methods achieved great processes and can fulfill the requirements of most scenarios, they still exhibit certain limitations. On the one hand, these methods usually develop the same mathematical model to indiscriminately extract image features, and rarely consider the inherent distinctiveness of different modality images, which limits the fusion performance improvement. On the other hand, the fusion rules or activity level measurement need to be manually designed. This strategy potentially compromises the objectivity and reliability of image fusion output, which is unsuitable for some complicated scenarios and subsequent decision-making applications.

In recent years, deep neural networks have experienced rapid adoption in the field of image fusion. Generally, the mainstream deep learning-based models include autoencoder (AE)-based [10], [11], convolutional neural network (CNN)-based [12], [13], Transformer-based [14], [15], and generative adversarial network (GAN)-based [16], [17] methods. AE-based methods employ the encoder-decoder framework to extract and reconstruct features, and design a fusion layer to integrate their respective features. Nevertheless, the fusion strategies are still hand-crafted. CNN-based methods usually concatenate source images in the input stage as an image-level framework or integrate features in the fusion stage to form a feature-level framework. Different to CNN, Transformer-based methods employ a self-attention mechanism to model the long-range dependencies, and achieve state-of-the-art (SOTA) performance. However, the above methods are non-generative fusion schemes, which cannot take advantage of strong generative ability. Image fusion as a generative task, GAN-based methods employ adversarial training to constrain the same distribution of fused output and source images. Nevertheless, the tradeoff between generator and discriminator is difficult to follow during training, which presents a challenge for achieving controlled generation. Moreover, unexplainable mechanism and mode collapse of GANs seriously affect the fusion quality.

Recently, denoising diffusion probabilistic models (DDPM) [18] have demonstrated remarkable advances in generating hopeful synthetic samples. Unlike the existing GAN-based methods, the generation process of DDPM is interpretable as it relies on denoising principles, which can effectively achieve controllable high-quality and high-fidelity generation. Furthermore, DDPM does not require discriminative constraints, thereby avoiding the common issues of unstable training and mode collapse often encountered by GANs. Specifically, Zhao et al. [19] formulated fusion task into an unconditional generation problem, and integrated the hierarchical Bayesian model in likelihood rectification. Yue et al. [20] constructed the multi-channel distribution based on diffusion model to extract complementary information for high color fidelity fusion tasks. Although these methods achieve surprising fusion performance, some drawbacks still need to be addressed. On the one hand, due to the posterior sampling procedure, their fusion models usually require extensive storage space and long inference times. On the other hand, these methods only leverage the generative capacity of diffusion mode while failing to consider the contextual interactions of multi-modality images, resulting in limited fusion performance.

To address these issues, we introduce a simple yet strong fusion baseline, namely diffusion model guided cross-attention learning network, termed as DMFuse. In the first training stage, to alleviate the strains on storage space and inference process, we directly compress the quadruple channels of diffusion UNet, and train a robust model using the MS-COCO dataset [21]. Because this dataset encompasses diverse object categories, abundant image data, and various visual scenarios, it aids in bolstering the generalization ability of the diffusion model for fusion tasks, even when model parameters are compressed. In the second training stage, instead of relying on mainstream convolution operations or self-attention mechanisms, we employ the pre-trained diffusion model as an autoencoder to generate the diffusion features, which can seamlessly transfer its high-quality generation ability to the subsequent fusion network. In addition, we develop a cross-attention interactive

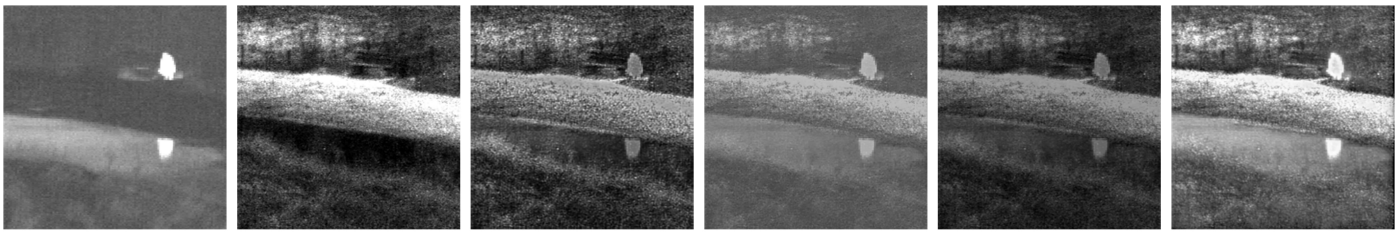


Figure 1. The comparative schematic diagram of the proposed model with U2Fusion [12], YDTR [15] and DDFM [19].

fusion module to aggregate the diffusion features of infrared and visible images, which can model the global dependencies from local contexts and improve the complementary characteristics of different modalities. Finally, a multi-level decoder network is proposed to progressively reconstruct the fused output.

To demonstrate the superiority of the proposed DMFuse, we compare it with the CNN-based method, i.e., U2Fusion [12], Transformer-based method, i.e., YDTR [15], and diffusion model-based method, i.e., DDFM [19]. Figure 1 illustrates a schematic diagram for comparison. U2Fusion and YDTR are non-generative schemes that focus on modeling local features and local-global dependencies, respectively. Although the fused results preserve visible details well, they fail to retain the infrared target brightness. DDFM formulates the fusion task into unconditional generation and samples a fusion image from the posterior distribution. However, the generated result still exhibits limited preservation of target brightness. In contrast, the proposed model can simultaneously enable rich detail preservation and considerable intensity control. In summary, the main contributions of our work are threefold.

- We introduce a novel diffusion model guided fusion baseline. The pre-trained diffusion model is employed as an encoder to provide a powerful distribution mapping, thereby grafting its generation ability for fusion tasks.
- We develop a cross-attention interactive fusion module to model the global dependencies from local diffusion features, thus effectively strengthening and integrating the complementary characteristics of different modalities.
- We train a more efficient and robust diffusion model with different strategies. Extensive experiments demonstrate that DMFuse achieves SOTA fusion performance as well as competitive operational efficiency.

The rest of this paper is schemed as follows. Section 2 mainly discusses the non-generative and generative fusion schemes. In Section 3, the framework of the proposed model is elaborated. In Section 4 and Section 5, experimental comparisons and relevant conclusions are given, respectively.

2 Related Work

This section provides an overview of the related work that is closely related to the proposed method. From a generative standpoint, we can roughly categorize the existing works as non-generative and generative fusion schemes.

2.1 Non-Generative Fusion Scheme

AE-based methods generally follow the traditional framework, and employ a pre-trained encoder-decoder network to extract and reconstruct features. For example, Li et al. developed DenseFuse [10] and NestFuse [11] where dense blocks and nest connections are introduced to enhance feature representation. Zhao et al. [22] presented AUIF in which the traditional optimization model was mapped to a trainable neural network by the algorithm unrolling. To improve fusion performance, Jian et al. elaborated SEDRFuse [23] and DDNSA [24] in which attention-based fusion strategies are employed to better strengthen the complementary characteristics of different modalities. However, these methods need to design the fusion strategies manually, restricting their practical applications.

CNN-based methods usually propose image-level or feature-level frameworks to implement unsupervised learning. Typically, Xu et al. [12] introduced U2Fusion, which concatenated source images as an input, and employed a pre-trained VGG-16 network to measure information preservation degree for supervising the similarity constraint. Li et al. [13] elaborated RFN-Nest, which proposed a two-stage training strategy to train the encoder-decoder framework and fusion network, respectively. They also presented

LRRNet [25], which formulated the fusion task as optimized decomposition and network learning problems. An et al. [26] introduced MRASFusion, which designed a residual attention fusion module for feature interactions. Chen et al. developed IVIFD [27] for a joint fusion and detection task. Zhu et al. [28] proposed MGRCFusion, which utilized a multi-scale group residual convolution module to exploit finer deep-level features.

Transformer-based methods mainly depend on the self-attention mechanism to model the global dependencies and maintain long-range context. Pang et al. [14] introduced SDTFusion, which employed dense Transformer blocks to extract the global features. Tang et al. presented YTDR [15] and DATFuse [29], which proposed a serial CNN-Transformer architecture to aggregate local and global features. Ma et al. [30] elaborated SwinFusion, which designed self-attention and cross-attention units to integrate intra- and inter-domain interactions. Tang et al. [31] developed a multi-branch network based on CNN and Transformer to extract the local and global information for multi-modality fusion. In addition, Liu et al. [32] introduced SegMiF, which proposed a multi-interactive framework for the joint tasks of fusion and segmentation.

The aforementioned methods tend to design efficient network structures [10], [11], [26], [28], novel fusion rules [23], [24], different training strategies [13], [22], [25], [27], long-range modeling [14], [15], [30], [31], and multi-task learning [12], [32]. The core is to employ convolutional or self-attention operations to discriminate model local, global, or joint features. However, due to the lack of ground truth and the fact that these methods are non-generative fusion schemes, the lack of in-depth exploration of generative models limits the potential fusion performance improvement.

2.2 Generative Fusion Scheme

GAN-based methods generally apply adversarial training to generate a fused image that follows the same distribution as the source images. Ma et al. [16] firstly devised FusionGAN, which employed a generator to obtain the fused image, and used a discriminator to determine whether the fused output has a similar distribution to source images. Meanwhile, they also introduced TarDAL [33], which designed a target-aware dual adversarial learning network for the joint problems of fusion and detection. Wang et al. presented ICAFusion [34], CrossFuse [35], and FreqGAN [36], which introduced attention mechanisms and frequency information to implement feature interaction and iterative optimization. These methods focus on the design of flexible networks, such as generator architecture [16], attention mechanism [34], [35], and multi-task learning [33]. However, the GAN-based methods suffer from unexplained mechanism, unstable training, and mode collapse, which adversely impacts the fusion quality.

Diffusion-based methods formulate fusion tasks as a conditional generation problem within the diffusion sampling framework, which can overcome the common problems of GANs. For example, Yue et al. [20] presented Dif-Fusion, which directly introduced the multi-channel data construction into a diffusion process, and achieved a fused output with high color fidelity. Zhao et al. [19] devised DDFM, where an unconditional generation module and a conditional likelihood rectification module are designed to deliver favorable results. These methods leverage the generative ability of diffusion mode, but present significant time-consuming issues in terms of storage space and inference processes, and do not take into account the contextual interactions. Different from them, the proposed model employs a more efficient and robust diffusion model to graft its high-quality generation ability for fusion tasks. Meanwhile, we design a cross-attention interactive fusion module to strengthen the complementary characteristics of different modalities. Therefore, the proposed model achieves superior fusion performance while requiring less computational costs.

3 Methodology

In this section, we elaborate on the overall workflow of the fusion baseline, including network overview, cross-attention interactive fusion module, and loss function.

3.1 Network Overview

As depicted in Figure 2(a), DMFuse consists of three core components, i.e., pre-trained diffusion model, multi-level decoder, and cross-attention interactive fusion module. Given the input infrared and visible images $I_0 = \{I_i, I_v\}$, the forward process of the diffusion model gradually adds Gaussian noise to the input image I_0, and generates noisy image $I_t = \{I_t^i, I_t^v\}$ and its distribution $P(I_t|I_{t-1})$ at timestep t.

After that, we employ the diffusion model encoder to extract multi-level diffusion features of infrared and visible images, termed as Φ_i^l and Φ_v^l, and fed them

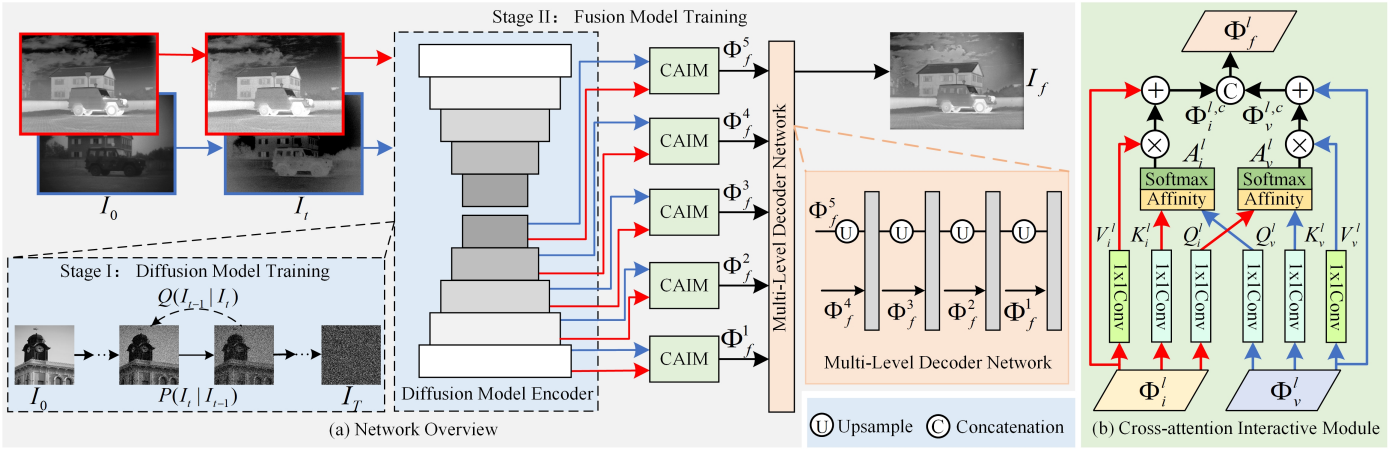


Figure 2. The overall workflow for the proposed model. The diffusion encoder is employed as autoencoder to extract the diffusion features from different modality images. And these features are fed into cross-attention interactive modules (CAIMs) to generate the fusion features. Finally, the fused output is reconstructed by a multi-level decoder network.

into cross-attention interactive fusion module (CAIM), which is shown in Figure 2(b), to generate the fusion features Φ_f^l. Finally, a multi-level decoder network is proposed to reconstruct the final fused outputs, which is formulated by Eq.(1).

$$I_f = C[\Phi_f^1, U(C[\Phi_f^2, U(C[\Phi_f^3, U(C[\Phi_f^4, U(\Phi_f^5)])])])] \tag{1}$$

where $C(\cdot)$ and $U(\cdot)$ denote the convolutional and upsampling operations. $[\cdot]$ indicates the channel concatenation. Next, we will describe the training process of the diffusion model.

3.2 Diffusion model encoder

The diffusion model implements the variational inference on a Markovian chain, which includes both forward and backward processes. In the forward process, Gaussian noise is incrementally added to the input image I_0 until it is fully destroyed within T timesteps. By using the reparameterization trick, the simplified distribution of noisy image I_t at each time step t can be directly derived from the input image I_0 sampling, which is formulated by Eq.(2).

$$P(I_t|I_0) = \mathcal{N}(I_t; \sqrt{\overline{\alpha}_t} I_0, (1-\overline{\alpha}_t)X) \tag{2}$$

where $\mathcal{N}$ is a Gaussian distribution, α_t denotes the variance schedule, and $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $t \in [1, T]$. X represents the standard normal distribution.

Technically, the forward process aims to degrade the image data into an isotropic Gaussian distribution by adding noise. On the contrary, the backward process attempts to eliminate the degradation by a denoising network. During the backward process, a series of denoising operations are performed on the noisy image I_t to obtain back I_{t-1}. The corresponding distribution of I_{t-1} given I_t can be formulated by Eq.(3).

$$Q(I_{t-1}|I_t) = \mathcal{N}(I_t; \mu_\theta(I_t, t), \sigma_t^2 X) \tag{3}$$

where $\mu_\theta(I_t, t)$ and σ_t^2 are the mean and standard deviation of $Q(I_{t-1}|I_t)$.

During the training phase, the noise $\varepsilon \sim \mathcal{N}(0, X)$ and the timestep $t \sim U(\{1, \cdots T\}$ are sampled from the standard normal distribution and the uniform distribution, respectively. The noisy image I_t and the timestep t are fed into the denoising network $\varepsilon_\theta(\cdot, \cdot)$, which is a UNet framework. A simple supervised loss can be formulated by Eq.(4).

$$L_{diff} = \left\| \varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha}_t} I_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, t) \right\|_2 \tag{4}$$

The diffusion model consists of a five-level U-Net framework, where the decoder backbone is subjected to randomly sampled noise levels to reconstruct the denoised diffusion features. Therefore, we employ the diffusion model as an encoder to extract multi-level diffusion features from noised infrared and visible images. The formulation is expressed by Eq.(5).

$$\{\Phi_i^l,\ \Phi_v^l\} = Dif\{I_t^i,\ I_t^v\} \tag{5}$$

where $Dif\{\cdot\}$ denotes the diffusion model encoder operation.

In particular, the diffusion model encoder is capable of generating more robust feature representations over the CNN encoder. Additionally, to accelerate inference process of the diffusion model, we compress the channel numbers of each layer to 1/4 of the original. A comprehensive discussion regarding the diffusion model encoder and its training strategies will be presented in the ablation study.

3.3 Cross-attention interactive fusion module

After training the diffusion model, we employ it as an encoder and freeze its parameters while proceeding to train the fusion network. The multi-level diffusion features are then utilized as input for the cross-attention interactive fusion modules, facilitating global interactions. Inspired by CCNet [37], we aggregate contextual dependencies together for all pixels in its criss-cross path. More importantly, we exchange the query features of different modalities to capture their interactive cross-attention maps, which effectively strengthens their complementary characteristics to promote better fusion performance.

As shown in Figure 2(b), given the diffusion features Φ_i^l and $\Phi_v^l \in R^{C\times H\times W}$, we first perform two convolution layers with 1×1 filters to achieve their query and key features, i.e., $\{Q_i^l, K_i^l\}$ and $\{Q_v^l, K_v^l\} \in R^{C'\times H\times W}$, where H and W represent the height and width of feature maps, and the channel C' is less than C for dimension reduction. After that, we exchange the feature maps Q_i^l and Q_v^l of different modalities and further generate their respective cross-attention maps A_i^l and $A_v^l \in R^{(H+W-1)\times(H\times W)}$ via *Affinity* opertions. Taking the infrared modality as an example, at the position n within the spatial dimension of infrared features K_i^l, we can achieve a vector $K_{i,n}^l$ from itself and a set $Q_{v,n}^l$ from visible features Q_v^l, which are in the same column or row with position n. Then, the *Affinity* opertions can be formulated by Eq.(6) and Eq.(7), respectively.

$$d_{i,m,n}^l = K_{i,n}^l Q_{v,m,n}^l \tag{6}$$

$$d_{v,m,n}^l = K_{v,n}^l Q_{i,m,n}^l \tag{7}$$

where $\{d_{i,m,n}^l, d_{v,m,n}^l\} \in \{D_i^l, D_v^l\}$ denote the degree of correlation between infrared and visible features and their reverse order, $\{Q_{i,m,n}^l, Q_{v,m,n}^l\} \in R^{C'}$ stand for the mth element of $Q_{i,n}^l$ and $Q_{v,n}^l$, $m = [1, \cdots, H+W-1]$, and $\{D_i^l, D_v^l\} \in R^{(H+W-1)\times(H\times W)}$. Then, we employ a softmax layer on D_i^l and D_v^l across the channel dimension to calcuate the cross-attention maps A_i^l and A_v^l, respectively.

Subsequently, another convolution layer with 1×1 filters is used for the diffusion features $\{\Phi_i^l, \Phi_v^l\}$ to generate $\{V_i^l, V_v^l\}$ for feature adaptation. Similarly, we can also obtain the vetors $\{V_{i,n}^l, V_{v,n}^l\} \in R^C$ and sets $\{V_{i,m,n}^l, V_{v,m,n}^l\} \in R^{(H+W-1)\times C}$ at their spatial position n. Thus, we apply an multiplication operation and a skip connection to collect the contextual information of different modalities, which are expressed by Eq.(8) and Eq.(9), respectively.

$$\Phi_i^{l,c} = \sum_{m=0}^{H+W-1} A_{i,m,n}^l V_{i,m,n}^l + \Phi_{i,n}^l \tag{8}$$

$$\Phi_v^{l,c} = \sum_{m=0}^{H+W-1} A_{v,m,n}^l V_{v,m,n}^l + \Phi_{v,n}^l \tag{9}$$

where $\Phi_i^{l,c}$ and $\Phi_v^{l,c}$ denote the global cross-attention features of infrared and visible modalities. Finally, we concatenate them to generate the fusion features Φ_f^l.

3.4 Loss function

To train the fusion model, we employ structural similarity (SSIM) loss, intensity loss, and gradient loss to supervise the network. Concretely, SSIM loss (L_{ssim}) is used to constrain the structural similarity between fused result I_f and source images I_i, I_v, which is defined by Eq.(10).

$$L_{ssim} = \omega_1(1 - ssim(I_f, I_i)) + \omega_2(1 - ssim(I_f, I_v)) \tag{10}$$

where $ssim(\cdot)$ denotes the structural similarity operation. ω_1 and ω_2 are set to 0.5.

Meanwhile, the intensity loss L_{int} is designed to maintain more valuable pixel intensity information from source images, and its formalization is expressed by Eq.(11).

$$L_{int} = \frac{1}{HW}\|I_f - mean(I_i, I_v)\|_1 \tag{11}$$

where $mean(\cdot)$ denotes the average operation.

Moreover, the gradient loss L_{grad} is proposed to transfer as many details as possible from different modalities, which is formulated by Eq.(12).

$$L_{grad} = \frac{1}{HW}\||\nabla I_f| - \max(|\nabla I_i|, |\nabla I_v|)\|_1 \tag{12}$$

where ∇ is the Sobel gradient operator. $\max(\cdot)$ and $\|\cdot\|_1$ stand for the maximum and L1-norm operations, respectively.

Finally, the total fusion loss can be expressed by Eq.(13).

$$L_{fusion} = \lambda_1 L_{ssim} + \lambda_2 L_{int} + \lambda_3 L_{grad} \tag{13}$$

where λ_1, λ_2 and λ_3 are the hyperparameters, which are used to balance the three losses.

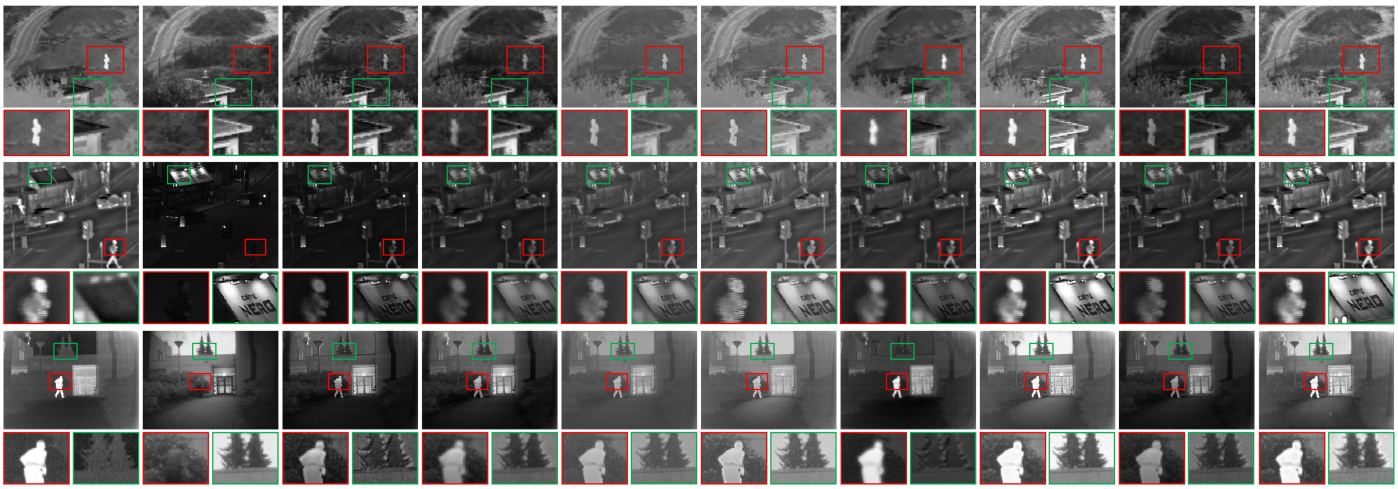


Figure 3. Visual descriptions of DMFuse with other SOTA competitors on the TNO benchmark.

4 Experimental Results and Analysis

This section introduces the correlative experimental configurations and comparative validations of fusion tasks and downstream applications. The ablation studies are also deeply discussed.

4.1 Experimental Configurations

In the training phase, we first train the diffusion model on the MS-COCO benchmark. This dataset includes more than 80000 complex scenario images. The training parameter settings are consistent with DDPM [18]. After that, we then train the fusion model on the TNO benchmark. To augment the training dataset, we take a sliding step of 12, crop the images into patches of size 256×256 and normalize their gray value range to [-1, 1]. This process yields a total of 18813 patch pairs for training. The batch size and number of epochs are set to 4 and 16, respectively. The model is optimized using the Adam optimizer. In the loss function, we empirically set λ_1, λ_2, and λ_3 to 1, 4, and 20. Additionally, the pre-trained diffusion model generates diffusion features at three different time steps, i.e., 5, 50, and 100. All experiments are conducted on a platform equipped with an NVIDIA GeForce GTX 3090, Intel I9-10850K, and 64 GB memory.

In the testing phase, we employ the TNO [1], M^3FD [2] and Harvard MIF [3] benchmarks, and select 25, 40 and 50 image pairs to evaluate the effectiveness and superiority of the proposed model. In addition, seven SOTA competitors, including the non-generative schemes, U2Fusion [12], RFN-Nest [13], YDTR [15], and DATFuse [29], the generative schemes, FusionGAN [16], Dif-Fusion [20], and DDFM [19], are selected to compare with the proposed model. Moreover, we also employ eight metrics, namely entropy (EN) [38], standard deviation (SD) [39], phase congruency (PC) [40], feature mutual information based on pixel (FMIp) [41], Qe [42], Qabf [43], multi-scale structural similarity (MS-SSIM) [44], and visual information fidelity (VIF) [45] for quantitative verification. In the follow-up experiments, the red bold and blue underline indicate the optimal and suboptimal values, respectively.

4.2 Results on TNO Benchmark

We first conduct experiments on the TNO benchmark to showcase the effectiveness of the proposed DMFuse. Three representative examples, namely *Nato_camp*, *Street*, and *Kaptein_1123*, are selected for subjective description, and their contrastive results are shown in Figure 3. The CNN-based methods, *i.e.*, U2Fusion and RFN-Nest, focus on modeling local features using image-level and feature-level frameworks, respectively. Although they manage to preserve visible details, they tend to lose brightness in the infrared targets. The Transformer-based methods, i.e., YDTR and DATFuse, attempt to integrate local and global features to achieve better visual effects. However, they still struggle to effectively control the brightness information. FusionGAN aims to retain target brightness but sacrifices visible detail information potentially due to unstable training. DDFM integrates inference solution and diffusion sampling within the same

[1][Online]. Available: https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029
[2][Online]. Available: https://github.com/dlut-dimt/TarDAL
[3][Online]. Available: http://www.med.harvard.edu/AANLIB/home.html.

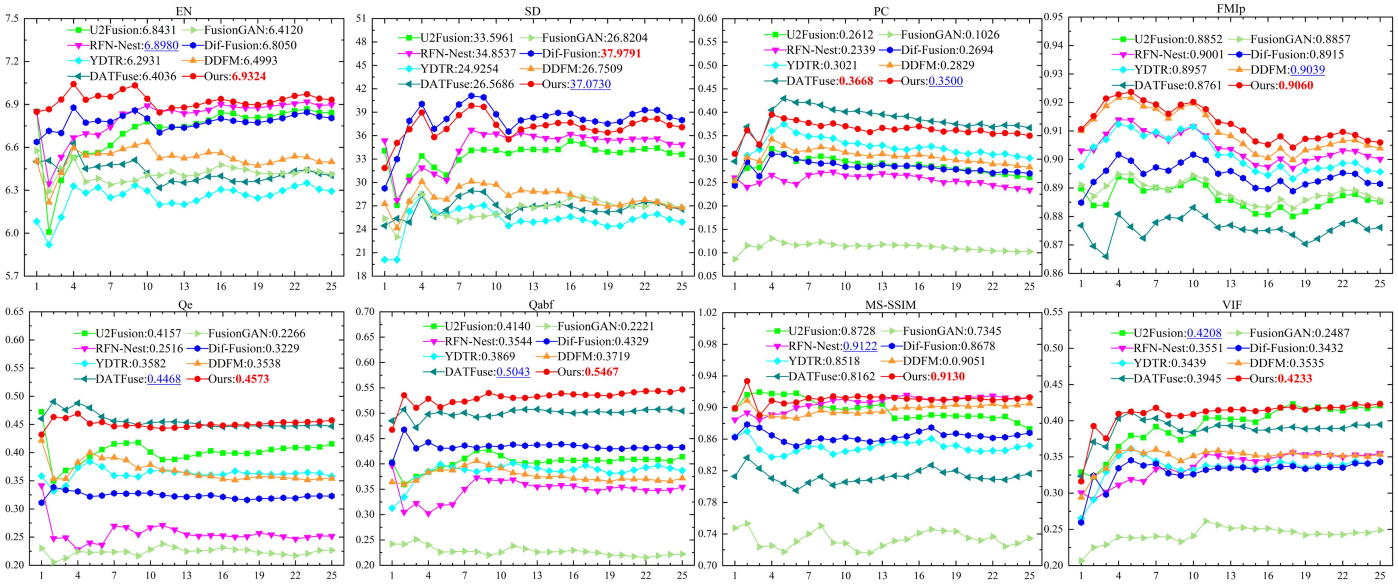


Figure 4. Quantitative comparisons of DMFuse with other SOTA competitors on the TNO benchmark.

Figure 5. Visual descriptions of DMFuse with other SOTA competitors on the M^{3}FD benchmark.

iterative framework to generate fusion images directly, but it fails to effectively combine thermal radiation information. Dif-Fusion constructs a multi-channel data distribution and yields similar results to the proposed model. In comparison, the proposed model effectively preserves rich details and control considerable intensity.

Subsequently, eight metrics previously mentioned are used for the quantitative evaluation of fusion performance, and the comparable results are presented in Figure 4. The proposed model is described by the red dotted line. Obviously, the proposed model demonstrates excellent performance across all metrics. The corresponding EN, FMIp, Qe, Qabf, MS-SSIM, VIF rank first, and SD, PC rank second, which follow behind Dif-Fusion and DATFuse, respectively. The optimal Qe, Qabf, and MS-SSIM indicate that the proposed model can transfer edge, gradient, and structural information into the fused results from source images. The optimal EN, FMIp, and suboptimal PC demonstrate that the proposed model can preserve significant details and meaningful information. The optimal VIF and suboptimal SD reveal that the proposed model has better visual performance and contrast definition. Quantitative experiments confirm its superiority, aligning with the above qualitative observations.

4.3 Results on M^{3}FD Benchmark

We further carry out experiments on the M^{3}FD benchmark, and compare the proposed model with

other competitors to verify its generalization ability. For the color image fusion, we first transfer the RGB visible image to the YCbCr color space, and return it after the Y channel is integrated with the infrared image. Figure 5 gives the subjective comparison results of three examples, namely 03878, 03989, and 00762. The proposed method offers significant advantages in terms of detail preservation and intensity control. For the salient pedestrian targets, the proposed model preserves high-brightness target characteristics and distinct contour edges. Meanwhile, for the background details, such as trees, windows, and handrails, it also gets the clearest detail description. In addition, Figure 6 describes the objective comparison results. The proposed model achieves the top ranking for all the metrics except for EN and SD, which are in arrears of Dif-Fusion. Both subjective and objective experiments demonstrate that the proposed model yields promising fusion performance and transcends other SOTA competitors.

4.4 Results on Harvard MIF Benchmark

In this section, we conduct experiments on the Harvard MIF benchmark to further verify the generalization of the proposed model. Figure 7 gives the subjective comparison results of three examples, namely MRI_CT_21, MRI_PET_32, and MRI_SPECT_48. Compared with other methods, the proposed model remains effectively the soft tissue texture information presented in MRI images and highlights the areas of high-density contrast enhancement in T images. Table. 1 presents the quantitative results of different fusion methods. Obviously, DMFuse obtains the optimal performance in terms of EN, SD PC, Qe, Qabf and VIF. The metrics FMIp and MS-SSIM rank second, which follow behind DDFM and Dif-Fusion, respectively. Both subjective and objective experiments demonstrate that the proposed model yields excellent performance in the medical image fusion tasks.

In summary, the above experiments on the TNO , M^3FD and Harvard MIF benchmarks confirm the superior performance and generalization ability of the proposed model for different lighting and object categories. The main reasons are twofold. On the one hand, we use the MS-COCO dataset to train the diffusion model for more stable performance. More importantly, we employ the diffusion model to guide the fusion network. The diffusion features fully exhibit a strong distribution mapping capacity, and provide extra feature details for fusion tasks. Therefore, the fused results preserve rich details from source images. On the other hand, the designed cross-attention interactive fusion module can effectively implement the global interactions of different modalities. Under the supervision of the loss function, the fusion images achieve better visual effects with high-brightness targets and unambiguous details. As a result, DMFuse makes the fusion image easy to distinguish foreground objects and background edges.

4.5 Downstream Application

In addition to fusion performance evaluation, we also explore the positive role of image fusion for downstream applications. Specifically, we analyze the effects of other visual tasks, such as object detection and semantic segmentation.

Image fusion for object detection: We first discuss how image fusion affects object detection performance. The experiments are implemented on the M^3FD benchmark, which contains 4200 images annotated with 33,603 objects, including six classes, i.e., People, Car, Bus, Motorcycle, Truck and Lamp. The YOLOv5 [46] network is used as the detection baseline, and mean average precision (mAP) is employed as the evaluation metric. Especially, mAP@0.5 represents the precision value at an intersection-over-union (IoU) threshold of 0.5, and mAP@[0.5:0.97] indicates the mean value at IoU thresholds of between 0.5 and 0.97, with steps of 0.05. For a fair comparison, we employ the detection model to source images and fused results.

Figure 8 presents the visual results of object detection. For the representative objects, such as People and Car, the proposed model achieves higher precision values than source images and other competitors, indicating that our fused results are more conducive to object detection tasks. Moreover, the objective comparison results are shown in Table 2. Almost all fusion methods yield good detection performance, and their mAP values are much better than those using only infrared or visible images. Notably, the proposed model outperforms other competitors in terms of mAP value, which has an improvement of 1.09% and 1.77% for mAP@0.5 and mAP@[0.5:0.97]. This indicates that the proposed model can fully discover unique information from different modalities, and offer effective complementary characteristics for the detector to achieve better performance.

Image fusion for semantic segmentation: We further evaluate the proposed DMFuse with other competitors

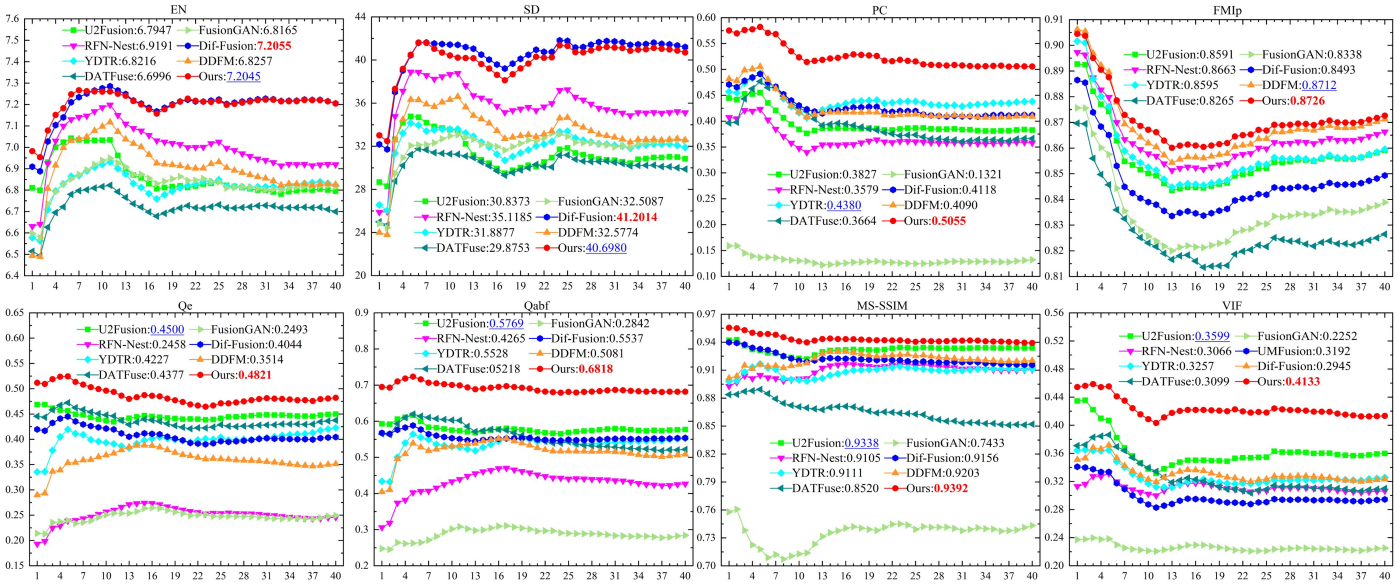


Figure 6. Quantitative comparisons of DMFuse with other SOTA competitors on the M^3FD benchmark.

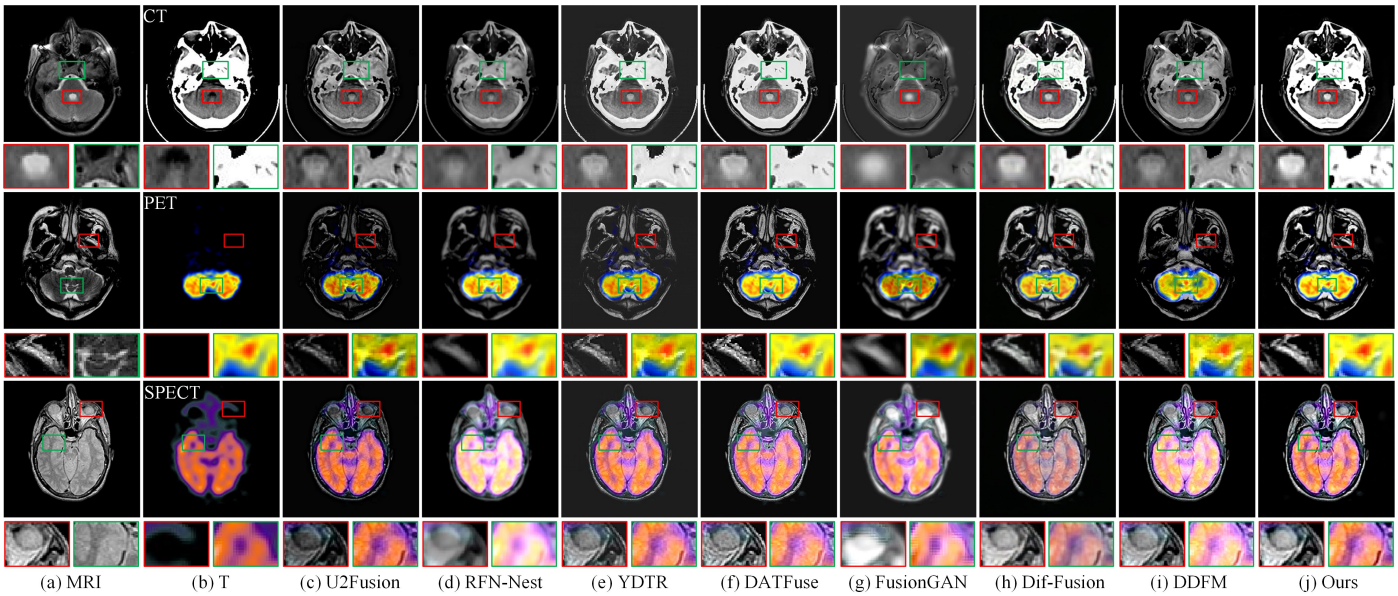


Figure 7. Visual descriptions of DMFuse with other SOTA competitors on the Harvard MIF benchmark.

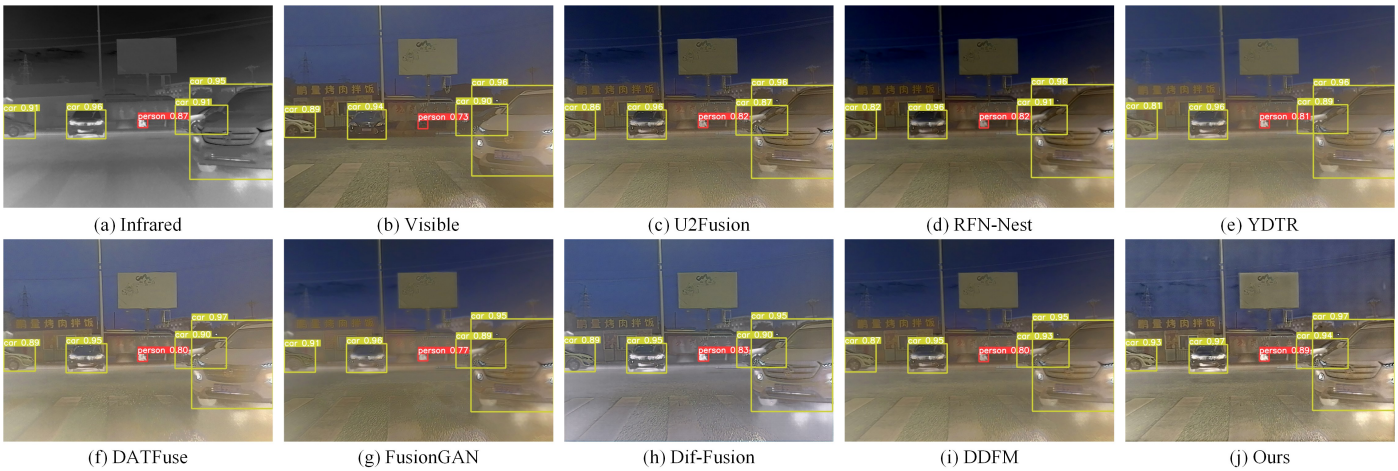


Figure 8. Qualitative object detection comparisons of source images and the fused results obtained by different methods.

Table 1. Quantitative comparisons of DMFuse with other SOTA competitors on the Harvard MIF benchmark.

Models	EN	SD	PC	FMIp	Qe	Qabf	MS-SSIM	VIF
U2Fusion [12]	3.7566	33.8763	0.3735	0.8579	0.3093	0.3776	0.8552	0.2489
RFN-Nest [13]	4.1351	56.6246	0.2396	0.8616	0.2229	0.1983	0.8928	0.2256
YDTR [15]	4.1527	37.6520	0.4553	0.8648	0.3990	0.4267	0.8811	0.2597
DATFuse [29]	4.2113	54.9562	0.4360	0.8531	0.5040	0.6113	0.9262	0.2605
FusionGAN [16]	4.2226	44.7076	0.1375	0.8496	0.2095	0.1662	0.8079	0.1708
Dif-Fusion [20]	4.7231	60.7802	0.4513	0.8660	0.4644	0.6354	**0.9559**	0.2994
DDFM [19]	3.8027	56.4941	0.4622	**0.8796**	0.4725	0.6363	0.9507	0.3288
Ours	**5.6969**	**61.8903**	**0.5438**	0.8754	0.5546	**0.7154**	0.9545	**0.3319**

Table 2. Quantitative object detection comparisons of different methods on the M^{3}FD benchmark.

Methods	mAP@0.5							mAP@[0.5:0.95]						
	Person	Car	Bus	Lamp	Motorcycle	Truck	All	Person	Car	Bus	Lamp	Motorcycle	Truck	All
Infrared	**0.783**	0.870	0.921	0.665	0.760	0.855	0.809	**0.551**	0.671	0.780	0.359	0.506	0.671	0.590
Visible	0.716	0.869	0.920	0.790	**0.790**	0.864	0.825	0.478	0.701	0.796	0.471	0.543	0.689	0.613
U2Fusion [12]	0.774	0.883	0.925	0.784	0.774	0.867	0.835	0.549	0.717	0.799	0.474	**0.547**	0.701	0.631
RFN-Nest [13]	0.772	0.881	0.924	0.790	0.775	0.865	0.835	0.544	0.716	0.798	0.467	0.541	0.700	0.628
YDTR [15]	0.768	0.885	0.925	0.781	0.766	0.859	0.831	0.546	0.714	0.800	0.473	0.539	0.700	0.629
DATFuse [29]	0.764	0.881	0.919	0.781	0.766	0.859	0.829	0.541	0.711	0.794	0.469	0.542	0.696	0.626
FusionGAN [16]	0.766	0.873	0.923	0.779	0.761	0.857	0.827	0.542	0.712	0.792	0.468	0.538	0.691	0.624
Dif-Fusion [20]	0.775	0.886	0.926	**0.796**	0.772	0.858	0.836	0.549	0.716	0.787	0.473	0.538	0.702	0.628
DDFM [19]	0.771	0.882	0.919	0.790	0.782	0.865	0.835	0.544	0.712	0.795	0.470	0.540	0.700	0.627
Ours	0.776	**0.887**	**0.927**	0.791	0.774	**0.875**	**0.838**	0.550	**0.719**	**0.806**	**0.475**	0.541	**0.710**	**0.634**

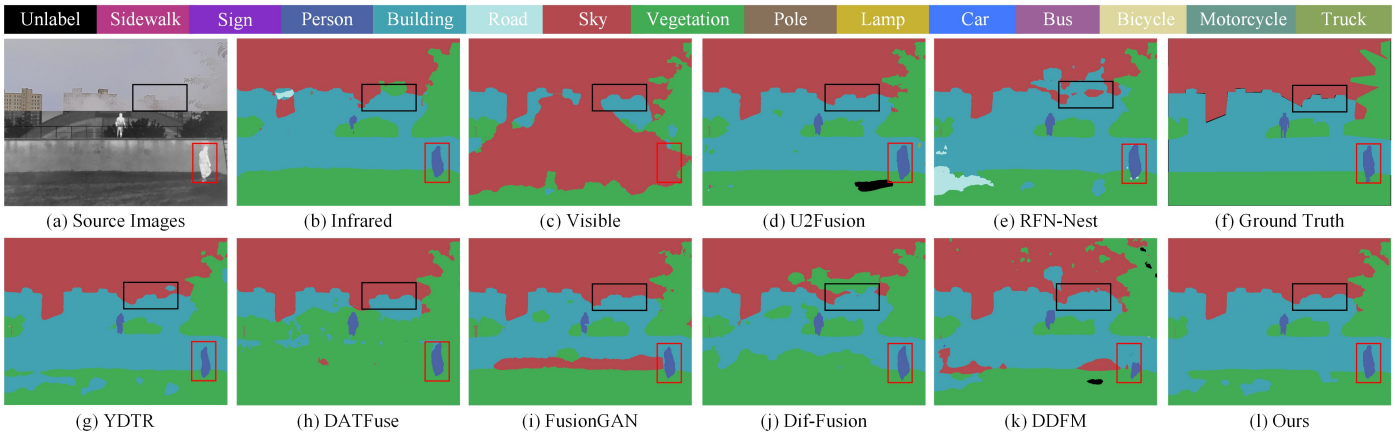


Figure 9. Qualitative semantic segmentation comparisons of DMFuse with other competitors on the FMB benchmark.

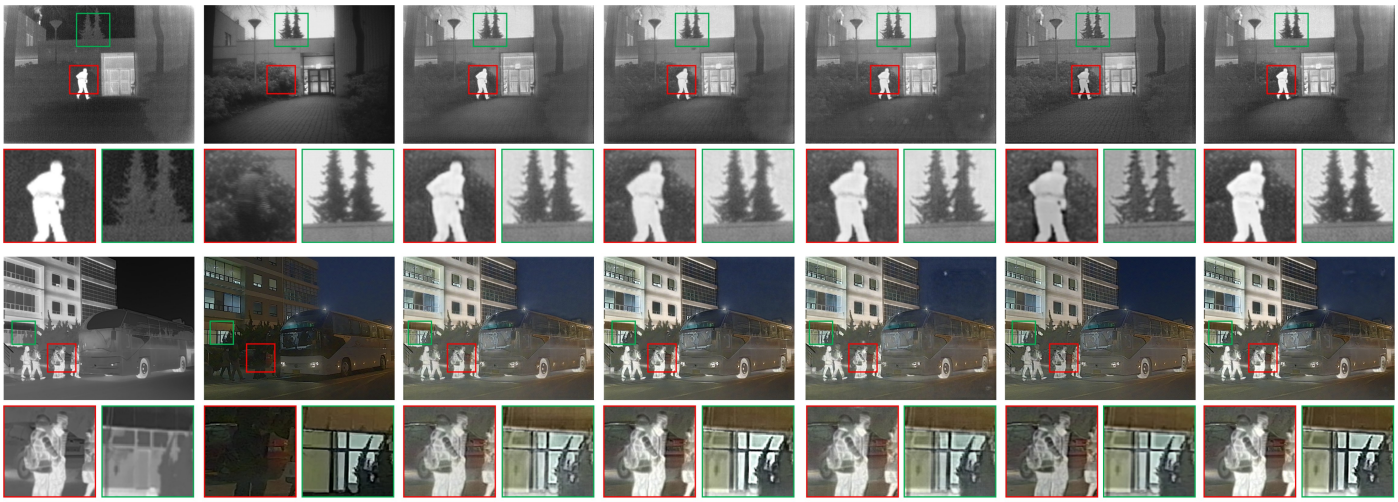
on the semantic segmentation task. A full-time multi-modality benchmark (FMB) [4] collected from the M^{3}FD benchmark is proposed for the segmentation baseline. The FMB dataset contains rich driving scenes under different lighting and weather conditions, and is labeled into fourteen categories. We select 1120 image pairs as the training set and verify the segmentation performance of different models on the 280 pairs. The relevant experimental configuration is derived from SegMiF [32]. The metrics, accuracy (ACC) and intersection-over-union (IoU) are employed for segmentation evaluation.

The qualitative semantic segmentation comparisons are depicted in Figure 9. For the representative objects and details, such as pedestrians and buildings, single-modality infrared and visible images cannot produce accurate classifications. However, the fusion methods improve the semantic segmentation performance to some extent. This indicates that

[4][Online]. Available: https://github.com/JinyuanLiu-CV/SegMiF

Table 3. Quantitative semantic segmentation comparisons of different methods on the FMB benchmark.

Methods	Road		Sidewalk		Lamp		Sign		Vegetation		Sky		Person		Pole		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Infrared	83.8	79.9	51.4	30.4	70.4	12.2	79.2	54.6	84.6	74.7	95.4	90.2	84.9	63.0	46.1	24.4	74.5	53.7
Visible	84.6	82.7	66.4	32.1	57.4	33.0	83.5	65.0	**93.0**	81.4	93.5	91.4	84.8	41.1	63.2	37.6	78.3	58.0
U2Fusion [12]	**91.1**	**85.3**	56.0	**39.6**	72.3	31.9	**86.5**	57.0	86.0	82.0	96.6	92.8	87.0	56.4	70.6	35.5	80.8	60.1
RFN-Nest [13]	84.7	76.3	62.1	36.3	**80.4**	24.9	77.8	68.3	91.9	82.2	96.7	93.9	85.6	60.8	70.1	39.2	81.2	60.2
YDTR [15]	83.9	81.3	72.4	33.5	61.6	27.8	73.3	66.4	89.7	84.0	95.6	93.9	83.4	58.5	**74.7**	39.0	79.4	60.6
DATFuse [29]	85.1	80.0	50.3	21.7	51.4	30.0	84.0	61.5	81.7	78.4	95.6	92.6	77.9	63.1	71.8	39.4	74.7	58.3
FusionGAN [16]	84.8	80.0	57.8	32.6	50.4	28.5	82.6	61.5	90.4	82.3	93.7	91.3	**89.2**	62.6	62.1	35.7	76.4	59.3
Dif-Fusion [20]	83.7	80.7	66.8	26.4	46.9	32.5	78.4	68.7	87.0	80.7	96.7	92.8	86.0	64.5	66.7	35.3	76.5	60.2
DDFM [19]	81.2	79.9	53.7	24.0	46.1	31.0	75.4	65.3	87.7	81.2	95.1	91.8	79.0	54.6	49.1	35.1	70.9	57.9
Ours	85.2	83.9	**73.0**	33.6	73.4	**43.6**	82.7	**70.3**	92.3	**85.6**	**97.3**	**94.5**	82.6	**67.5**	67.2	**48.2**	**81.7**	**65.9**



(a) Infrared (b) Visible (c) TNO (d) M^3 FD (e) w/o CAIM (f) w/o Dif (g) Ours

Figure 10. Visual comparisons of ablation experiments for two examples selected from the TNO and M^{3}FD benchmarks.

the complementary characteristics of image fusion facilitate the segmentation accuracy. More importantly, the proposed model effectively classifies objects and scenes with high accuracy, which is closest to ground truth. Table 3 reports the quantitative semantic segmentation comparisons. The numerical results demonstrate the proposed model is ahead of other SOTA competitors in terms of mACC and mIoU. In short, the proposed model can exploit and strengthen the complementary information of different modalities, which generates a positive effect on semantic segmentation.

4.6 Ablation Study

This section presents several specialized designs incorporated into the proposed DMFuse, and their effectiveness is evaluated through ablation experiments that focus on the model architecture and training strategy. The qualitative and quantitative comparisons are also presented in this section.

Training on Different Datasets: To assess the generalization performance of the diffusion model, we train it on the different datasets, including TNO, M^{3}FD, and the proposed MS-COCO. From the results of Figure 10 (c) and (d), the fusion images of TNO and M^{3}FD trained models exist in detail confusion and color degradation to a certain extent. The quantitative verification is compared in Table 4. A typical phenomenon is that a fusion model trained by a certain dataset maintains superior performance on the corresponding testing. Overall, the proposed method achieves more stable and outstanding performance on different testing datasets.

Channel in Diffusion UNet: We compress the channel numbers of diffusion UNet at each layer to 1/4 in our fusion model, and compare it with other competitive models, i.e., original parameters, 1/2, and 1/8. Noting that we omit the qualitative descriptions because their results are similar. Table

Table 4. Quantitative validations of different training datasets.

Testing Datasets	Training Datasets	EN	SD	PC	FMIp	Qe	Qabf	MS-SSIM	VIF
TNO Benchmark	TNO	6.8466	35.7474	0.3086	0.9026	0.4073	0.5009	0.9090	0.4154
	M^3FD	6.8466	34.0896	0.3032	0.9002	0.3936	0.4767	**0.9156**	0.3901
	MS-COCO (Ours)	**6.9324**	**37.0730**	**0.3500**	**0.9060**	**0.4573**	**0.5467**	0.9130	**0.4233**
M^3FD Benchmark	TNO	7.0188	36.4068	0.2798	0.8538	0.2723	0.4244	0.8990	0.2786
	M^3FD	7.1955	40.2199	0.3149	0.8487	0.3697	0.5227	0.9195	0.3068
	MS-COCO (Ours)	**7.2045**	**40.6980**	**0.5056**	**0.8726**	**0.4821**	**0.6818**	**0.9392**	**0.4133**

Table 5. Quantitative validations of different channels on the TNO benchmark.

Metrics	EN	SD	PC	FMIp	Qe	Qabf	MS-SSIM	VIF	Params(M)	FLOPs(G)	Time(s)
Original	6.9135	**37.6477**	**0.3845**	**0.9106**	**0.4861**	**0.5898**	**0.9150**	**0.4336**	392.724	1516.136	74.110
1/2	6.9150	37.1946	0.3738	0.9084	0.4794	0.5754	0.9125	0.4296	98.680	382.052	6.403
1/4(Ours)	6.9324	37.0730	0.3500	0.9060	0.4573	0.5467	0.9130	0.4233	24.967	106.584	2.624
1/8	**6.9402**	36.9426	0.2405	0.8899	0.3849	0.4181	0.9036	0.3786	**6.433**	**35.967**	**2.163**

Table 6. Quantitative validations of component effectiveness.

Models	EN	SD	PC	FMIp	Qe	Qabf	MS-SSIM	VIF
w/o Dif	6.8480	35.1861	0.3196	0.8975	**0.4735**	0.4862	0.8830	0.4228
w/o CAIM	6.8574	35.9839	0.3155	0.8886	0.3477	0.4902	0.8985	0.3439
Ours	**6.9324**	**37.0730**	**0.3500**	**0.9060**	0.4573	**0.5467**	**0.9130**	**0.4233**

5 shows the quantitative validations on the TNO benchmark. It can be observed that the fusion performance decreases with the reduction in channel numbers, while the model parameters and operation efficiency exhibit an opposite trend. When the channel parameter is reduced to 1/8, the performance becomes comparable to other fusion methods, such as Dif-Fusion and DDFM. In conclusion, the proposed model suggests adopting 1/4 channel parameters to achieve a better balance between fusion performance and computational efficiency.

Verification of Each Component: We employ the diffusion model to extract generative features and develop a cross-attention interactive fusion module to perform the global interactions. To verify their effectiveness, we propose an UNet-style CNN encoder to replace the diffusion model encoder and utilize addition operation instead of CAIM, respectively. As shown in Figure 10 (e) and (f), the fusion images without the diffusion model, termed w/o Dif, lose some target brightness and meaningful details, while the fused results without CAIM, termed w/o CAIM, have limited visual effects. Meanwhile, we visualize the feature maps of diffusion model encoder and CNN encoder (referred to as w/o Dif) in Figure 11. The diffusion features (the first row) demonstrate obvious advantages over CNN features (the second row) in the characterization of infrared salient targets and visible typical details. In addition, the quantitative results, as shown in Table 6, indicate that the proposed model achieves all the optimal values except for Qe, which is behind w/o Dif. The experiments prove that both diffusion model and CAIM are beneficial to fusion performance improvement.

Table 7. The computational efficiency comparisons.

Methods	Params.(M)	FLOPs(G)	Time(s)	
			TNO	M^3FD
U2Fusion [12]	0.659	43.17	1.722	4.646
RFN-Nest [13]	7.524	111.1	0.235	0.864
YDTR [15]	0.107	20.58	0.201	0.771
DATFuse [29]	**0.011**	**1.185**	**0.019**	**0.047**
FusionGAN [16]	1.314	57.09	0.513	0.988
Dif-Fusion [20]	434.2	726.1	4.820	17.21
DDFM [19]	988.3	2946	59.18	162.1
Ours	24.96	106.6	2.624	5.342

4.7 Efficiency Comparison

We also conduct experiments to evaluate the operational efficiency of different methods, including training parameters (Params.), floating-point operations per second (FLOPs), and runtime (Time). Table 7 presents their computational complexity. Note that the computation of FLOPs is implemented by a testing image with the size of 256×256. Compared

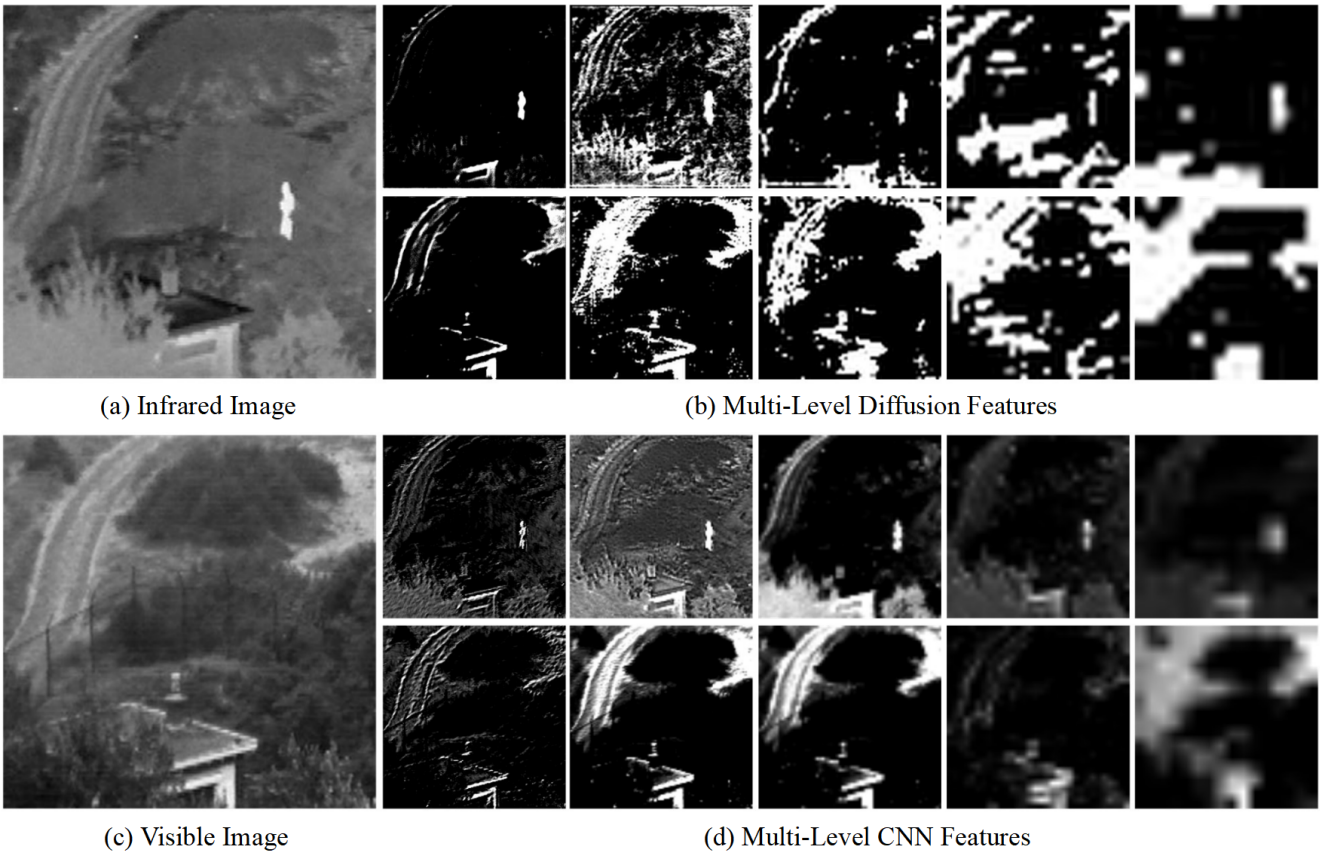


Figure 11. The visualization maps of different encoders.

with the diffusion-based methods, the non-generative fusion schemes, including U2Fusion, RFN-Nest, YDTR, DATFuse, and the GAN-based method, i.e., FusionGAN, have a significant advantage in terms of training parameters, FLOPs, and runtime. The main reason is that the diffusion model requires many iteration steps and consumes massive computational resources. However, since we train a more efficient model by compressing quadruple channels of diffusion UNet, the proposed model has higher operational efficiency than Dif-Fusion and DDFM, indicating the effectiveness of model training.

5 Discussion

The diffusion model showcases powerful generative capabilities and has manifested outstanding performance in the domain of image fusion. Nevertheless, its computational inefficiency constitutes a significant challenge because of the large quantity of iterative steps and the complexity of the calculations. These factors lead to a slow diffusion process, which restricts its applicability in scenarios demanding low computing resources. In future works, we aim to tackle these challenges by exploring optimization strategies such as sampling optimization [48] to reduce the number of iteration steps and latent space transformation [49] to streamline computations. These efforts will concentrate on enhancing computational efficiency while maintaining or improving the quality of the fused results.

6 Conclusion

This paper presents DMFuse, a novel diffusion model-guided cross-attention learning network, designed for infrared and visible image fusion. Unlike existing methods, the proposed model involves training a lightweight diffusion model to serve as an autoencoder, effectively integrating its high-quality generative capability into the fusion tasks. Moreover, we develop a cross-attention interactive fusion module that facilitates global interactions, strengthening the complementary characteristics of different modalities. We evaluate the performance of DMFuse against seven SOTA methods on TNO, M^3FD and Harvard MIF benchmarks. The experimental results validate the proposed model achieves predominant fusion performance and competitive computational efficiency. Furthermore, DMFuse exhibits positive implications for downstream applications, including object detection and semantic segmentation. In future

work, we will explore the integration of diffusion models with large language models (LLMs) [50], introducing text descriptions as a semantic guide to further enhance the quality of the fused images.

Data Availability Statement

The code and data supporting this study are publicly available on GitHub at the following link: https://github.com/Zhishe-Wang/DMFuse.

Funding

This work was supported in part by the Fundamental Research Program of Shanxi Province under Grant 202203021221144, and the Patent Transformation Program of Shanxi Province under Grant 202405012.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Liu, J., Wang, J., Huang, N., Zhang, Q., & Han, J. (2022). Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10), 7226-7240. [CrossRef]

[2] Wang, J., Song, K., Bao, Y., Huang, L., & Yan, Y. (2021). CGFNet: Cross-guided fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5), 2949-2961. [CrossRef]

[3] Wang, Y., Wei, X., Tang, X., Yu, K., & Luo, L. (2023). RGBT tracking using randomly projected CNN features. *Expert Systems with Applications*, 223, 119865. [CrossRef]

[4] Chen, J., Li, X., Luo, L., Mei, X., & Ma, J. (2020). Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Information Sciences*, 508, 64-78. [CrossRef]

[5] Li, H., Wu, X. J., & Kittler, J. (2020). MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29, 4733-4746. [CrossRef]

[6] Kong, W., Lei, Y., & Zhao, H. (2014). Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. *Infrared Physics & Technology*, 67, 161-172. [CrossRef]

[7] Ma, C., Nie, R., Ding, H., Cao, J., & Mei, J. (2023). A fractional-order variation with a novel norm to fuse infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, 72, 1-12. [CrossRef]

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

[9] Zhao, Z., Xu, S., Zhang, C., Liu, J., & Zhang, J. (2020). Bayesian fusion for infrared and visible images. *Signal Processing*, 177, 107734. [CrossRef]

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

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

[12] Xu, H., Ma, J., Jiang, J., Guo, X., & Ling, H. (2020). U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1), 502-518. [CrossRef]

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

[14] Pang, S., Huo, H., Liu, X., Zheng, B., & Li, J. (2024). SDTFusion: A split-head dense transformer based network for infrared and visible image fusion. *Infrared Physics & Technology*, 138, 105209. [CrossRef]

[15] Tang, W., He, F., & Liu, Y. (2022). YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25, 5413-5428. [CrossRef]

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

[17] Ma, J., Zhang, H., Shao, Z., Liang, P., & Xu, H. (2020). GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70, 1-14. [CrossRef]

[18] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.

[19] Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., ... & 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). [CrossRef]

[20] Yue, J., Fang, L., Xia, S., Deng, Y., & Ma, J. (2023). Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. *IEEE*

Transactions on Image Processing. [CrossRef]

[21] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

[22] Zhao, Z., Xu, S., Zhang, J., Liang, C., Zhang, C., & Liu, J. (2021). Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3), 1186-1196. [CrossRef]

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

[24] Jian, L., Rayhana, R., Ma, L., Wu, S., Liu, Z., & Jiang, H. (2021). Infrared and visible image fusion based on deep decomposition network and saliency analysis. *IEEE Transactions on Multimedia*, 24, 3314-3326. [CrossRef]

[25] Li, H., Xu, T., Wu, X. J., Lu, J., & Kittler, J. (2023). Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9), 11040-11052. [CrossRef]

[26] An, R., Liu, G., Qian, Y., Xing, M., & Tang, H. (2024). MRASFusion: A multi-scale residual attention infrared and visible image fusion network based on semantic segmentation guidance. *Infrared Physics & Technology*, 139, 105343. [CrossRef]

[27] Chen, B., Luo, S., Wu, H., Chen, M., & He, C. (2024). Infrared and visible image fusion and detection based on interactive training strategy and feature filter extraction module. *Optics & Laser Technology*, 179, 111383. [CrossRef]

[28] Zhu, P., Yin, Y., & Zhou, X. (2025). MGRCFusion: An infrared and visible image fusion network based on multi-scale group residual convolution. *Optics & Laser Technology*, 180, 111576. [CrossRef]

[29] Tang, W., He, F., Liu, Y., Duan, Y., & Si, T. (2023). DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7), 3159-3172. [CrossRef]

[30] Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7), 1200-1217. [CrossRef]

[31] Tang, W., He, F., & Liu, Y. (2023). TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation. *Pattern Recognition*, 137, 109295. [CrossRef]

[32] Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., ... & Fan, X. (2023). Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8115-8124). [CrossRef]

[33] 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). [CrossRef]

[34] Wang, Z., Shao, W., Chen, Y., Xu, J., & Zhang, X. (2022). Infrared and visible image fusion via interactive compensatory attention adversarial learning. *IEEE Transactions on Multimedia*, 25, 7800-7813. [CrossRef]

[35] Wang, Z., Shao, W., Chen, Y., Xu, J., & Zhang, L. (2023). A cross-scale iterative attentional adversarial fusion network for infrared and visible images. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8), 3677-3688. [CrossRef]

[36] Wang, Z., Zhang, Z., Qi, W., Yang, F., & Xu, J. (2024). FreqGAN: Infrared and Visible Image Fusion via Unified Frequency Adversarial Learning. *IEEE Transactions on Circuits and Systems for Video Technology*. [CrossRef]

[37] Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., & Huang, T. S. (2023). CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6), 6896-6908. [CrossRef]

[38] Roberts, J. W., Van Aardt, J. A., & Ahmed, F. B. (2008). Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1), 023522. [CrossRef]

[39] Rao, Y. J. (1997). In-fibre Bragg grating sensors. *Measurement science and technology*, 8(4), 355. [CrossRef]

[40] Liu, Z., Forsyth, D. S., & Laganière, R. (2008). A feature-based metric for the quantitative evaluation of pixel-level image fusion. *Computer Vision and Image Understanding*, 109(1), 56-68. [CrossRef]

[41] Haghighat, M., & Razian, M. A. (2014, October). Fast-FMI: Non-reference image fusion metric. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-3). IEEE. [CrossRef]

[42] Piella, G., & Heijmans, H. (2003, September). A new quality metric for image fusion. In *Proceedings 2003 international conference on image processing* (Cat. No. 03CH37429) (Vol. 3, pp. III-173). IEEE. [CrossRef]

[43] Xydeas, C. S., & Petrovic, V. (2000). Objective image fusion performance measure. *Electronics letters*, 36(4), 308-309.

[44] Ma, K., Zeng, K., & Wang, Z. (2015). Perceptual quality assessment for multi-exposure image fusion. *IEEE*

Transactions on Image Processing, 24(11), 3345-3356. [CrossRef]

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

[46] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 779–788).

[47] Li, S., van de Weijer, J., Khan, F., Liu, T., Li, L., Yang, S., ... & Cheng, M. M. (2023). Faster diffusion: Rethinking the role of the encoder for diffusion model inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[48] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695). [CrossRef]

[49] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Wuqiang Qi received the B.S degree in software engineering from Taiyuan University of Technology, Taiyuan, China, in 2021. He is currently pursuing the M.S. degree in optical engineering at Taiyuan University of Science and Technology, Taiyuan, China. His current research interests include image fusion and deep learning. (qiwq@stu.tyust.edu.cn)

Zhuoqun Zhang received the B.S degree in electrical engineering and automation from Tongji Zhejiang College, Jiaxing, China, in 2022. He is currently pursuing the M.S. degree in optoelectronic information engineering at Taiyuan University of Science and Technology, Taiyuan, China. His current research interests include image fusion and deep learning. (Email: zhangzhq@stu.tyust.edu.cn)

Zhishe Wang (Member, IEEE) received the B.S degree in automation from North China Institute of Technology, Taiyuan, China, in 2002. He received the M.S. and Ph.D. degree in signal and information processing from the North University of China, Taiyuan, China, in 2007 and 2015. He is currently a professor with Taiyuan University of Science and Technology. His current research interests include computer vision, pattern recognition and machine learning.(Email: wangzs@tyust.edu.cn)