DMNet: A Dense Multiscale Feature Extraction Network With Two-Stage Training for Infrared-Visible Image Fusion

Chengyi Pan, Qian Jiang, *Member, IEEE*, Huangqimei Zheng, Hongyue Huang, Xin Jin, *Senior Member, IEEE*, Keqin Li, *Fellow, IEEE*, and Wei Zhou, *Member, IEEE*

***Abstract*—With the increasing need for intelligent and secure multimedia systems, infrared and visible image fusion (IVIF) has garnered a lot of attention due to its ability to overcome the limitations of a single sensor and integrate unique information from different modalities. However, it is common to overlook how the spatial frequency information of visible and infrared images differs. A less thorough feature extraction may result from many approaches' inability to reconcile the extraction of both global and local information. To solve the aforementioned difficulties, we propose a dense multiscale fusion network DMNet. Through a dual-stream collaborative feature decoupling, the proposed network optimizes both the encoder–decoder network and the diffusion model to extract multimodal information more comprehensively. Specifically, the three-stage progressive encoder sequentially integrates dense transformer block (DTB) and dense invertible neural network block (DIB) to achieve global feature extraction and multimodal feature decoupling. Our proposed channel and spatial attention block (CSAB) selectively focuses on the important feature maps to better capture the critical information. Additionally, multiscale latent features are extracted by the diffusion module (DM) to enhance the representation of cross-modal latent features. As demonstrated by extensive experiments, DMNet outperforms representative state-of-the-art methods. Furthermore, we conduct sufficient ablation experiments to validate each module's effectiveness, and we demonstrate that DMNet can enhance downstream infrared-visible object detection performance. Our fused results and code will be accessible at https://github.com/Pancy9476/DMNet.**

***Index Terms*—Attention mechanism, dense network, diffusion model, image fusion, invertible neural network, transformer.**

I. Introduction

MULTIMODAL image fusion is widely needed in multimedia products. Due to hardware technological restrictions, infrared (IR) images are usually visually blurred, have low signal-to-noise ratio, low contrast, and low resolution [1], while visible light (VIS) images are subject to variations in weather changes, illumination, and other factors. Fusing IR and VIS pictures helps lessen a single sensor's flaws and maximize the advantages of both sensors, thus significantly improving image quality. Infrared-visible image fusion technology is extensively applied in various fields such as transportation, medicine, object detection, military actions, and agriculture [2], [3].

In recent years, many researchers have explored the image fusion field through deep learning methods [4], [5] to improve the fusion effect. Over the past decades, image fusion technologies have evolved from traditional signal processing-based methods (e.g., multiscale transforms, principal component analysis, and sparse representation) [6] to learning-based approaches [7]. Traditional methods rely heavily on handcrafted features and fail to fully adapt to complex image distributions. In contrast, deep-learning-based methods offer data-driven capabilities to automatically extract semantic features, significantly improving fusion performance. Particularly, feature fusion plays a pivotal role in determining the quality of the fused image, as it directly affects how complementary information from different modalities is integrated. Three commonly used networks based on CNN [8], autoencoder (AE) [9], and GAN [10] have been widely adopted.

Due to the differences in the spatial frequency (SF) features of the visible and infrared images, it may lead to a uniform luminance bias in the images, as shown in Fig. 1(a) and (b). In Fig. 1(a), the infrared image clearly preserves target details such as the person's outline and cloud contours, even under low-light and occluded conditions. In contrast, Fig. 1(b) shows that the corresponding visible image suffers from significant visual degradation due to poor illumination and fog. This SF mismatch becomes more pronounced after fusion and may degrade image quality if not handled appropriately. Infrared images typically have lower spatial frequencies, while VIS images have higher frequencies [11]. However, existing fusion methods usually suffer from the problem of unreasonable weight setting of different source images. In some methods, the proportion of VIS images in the fused result is too large,

Received 26 May 2025; revised 15 June 2025 and 8 July 2025; accepted 10 July 2025. Date of publication 16 July 2025; date of current version 25 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62261060; in part by the Yunnan Fundamental Research Projects under Grant 202301AW070007, Grant 202301AU070210, and Grant 202401AT070470; in part by the Yunnan Province Expert Workstations under Grant 202305AF150078; and in part by the Xingdian Talent Project in Yunnan Province of China. *(Corresponding author: Qian Jiang.)*

Chengyi Pan, Qian Jiang, Huangqimei Zheng, Hongyue Huang, Xin Jin, and Wei Zhou are with the School of Software, Yunnan University, Kunming 650000, China (e-mail: panchengyi@stu.ynu.edu.cn; jiangqian_1221@163.com; zhenghuangqimei@stu.ynu.edu.cn; huang.hongyue@ynu.edu.cn; xinxin_jin@163.com; zwei@ynu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/JIOT.2025.3589604

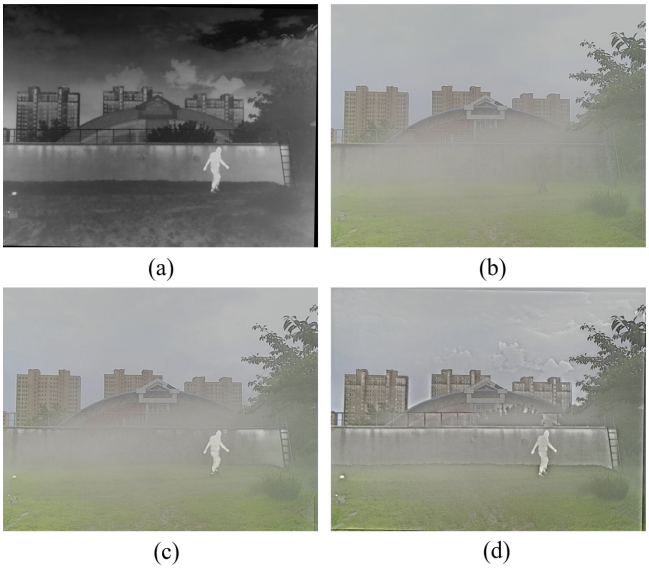

Fig. 1. Schematic illustration of IVIF on the 00361 image pair from the M3FD dataset. (a) Infrared. (b) Visible. (c) Diff-IF (INFFUS 2024). (d) Ours.

resulting in a wider brightness range of the final fused image, which reduces the visual clarity and detail expression of the image. As shown in Fig. 1(c), the enclosure behind the fog is obscured and the cloud information in the IR image is lost. Effectively fusing these two types of images requires the ability to extract features from different scales.

To address the above problems, we develop a fusion network called dense multiscale feature extraction (DMNet) for extracting features at different frequencies and balancing the weights of the source images to enhance the quality of the fused images. As seen in Fig. 1(d), our method not only effectively preserves the rich texture information from the source images, but also retains the target information from the infrared image. Unlike traditional simple cascade architectures, we innovatively establish a bidirectional collaborative mechanism between the encoder–decoder network and the diffusion model. This mechanism enables comprehensive cross-modal feature extraction through an improved dense feature enhancement module and a channel-spatial attention-guided block. The encoder adopts a three-stage progressive feature abstraction architecture, where the dense transformer block (DTB) captures long-range dependencies, while the dense invertible neural network block (DIB) facilitates cross-modal feature decoupling through nonlinear mapping. Additionally, our proposed channel and spatial attention block (CSAB) module improves the model's capacity to extract critical information by selectively focusing on important feature maps through the joint channel and spatial attention mechanism. The diffusion modelis particularly suitable for fusing visible and infrared images because the U-Net structure in its denoising network can extract multiscale features. In addition, we train the diffusion module by gradually adding and removing noise. The procedure forces the model to concentrate on extracting information from images at different scales. Since the training process exposes the model to various noise levels, this makes the model robust to noise in the input images.

Our proposed DMNet is collaboratively designed based on theoretical motivations, skillfully integrating the diffusion model, Transformer, invertible neural network, and attention mechanism to achieve unified optimization of information compression, modality decoupling, scale enhancement, and detail refinement. Our high-quality fused images not only retain the critical information from multisource images more accurately, but also significantly enhance the performance of downstream tasks. In applications such as autonomous driving [12], object detection [13], and intelligent surveillance [14], high-quality fused images can improve system recognition accuracy and decision-making capabilities. Our contribution is summarized as follows.

1) Considering the varying spatial frequencies of images, we suggest an innovative network DMNet to extract features at various spatial scales. On five popular datasets, our method improves the quality of fused images more than the state-of-the-art methods.
2) We innovatively employ the diffusion model as a feature extraction module for infrared and visible image fusion (IVIF) tasks. It highlights the multiscale features, crucial important to improve the fusion performance.
3) Our proposed DTB and DIB have significant advantages in multiscale feature extraction. The DTB utilizes a self-attentive mechanism to strengthen the capture of both local and global information, while the DIB through invertible mapping retains more image details and information. The combination of these two modules significantly improves the quality of fused images.
4) We propose the CSAB to generate weighted attention maps from channel and spatial dimensions. The CSAB module selectively focuses on important feature maps and suppresses unimportant features, thus enhancing the model's ability to capture critical information.

II. Related Work

This section presents related work and background material pertinent to the methodology presented in this work, including deep-learning-based methods, diffusion models and INN modules.

A. Deep-Learning-Based Methods

Deep learning networks, such as CNN, GAN, and AEs, are commonly used to solve infrared-visible image fusion problems. Compared to traditional methods, these networks have significant advantages in terms of fast fusion speed and clear fused images. LP-CNN [15] pioneered the use of CNN for image fusion, while IFCNN [8] eliminates manually-designed fusion rules. However, single CNN-based fusion architectures still have some limitations, such as insufficient feature extraction due to fewer convolutional layers, and pooling operations that may lead to loss of advanced features. To address these shortcomings, ReCoNet [16] introduces a deformation module and an attention mechanism, effectively enhancing robustness in misaligned scenarios. U2Fusion [17], on the other hand, builds upon DenseNet and further proposes a feature measurement framework, achieving multitask

fusion through adaptive information evaluation. Nevertheless, balancing local feature extraction and global context modeling remains a key challenge for CNN-based methods.

The AE framework offers a novel solution by leveraging an encoder–decoder structure to achieve feature disentanglement and reconstruction, making it another mainstream approach. The classical DenseFuse [9] enhances feature transmission using dense blocks, with its encoder extracting multiscale features through dense connections. CrossFuse [18] adopts a two-stage training (TST) strategy, utilizing a cross-attention mechanism to strengthen feature interactions within the encoder, thereby reducing redundant information while preserving key details. MUFusion [19] designs a memory unit architecture, where intermediate fusion results are used to self-evolve and optimize network parameters.

Since FusionGAN [20] first introduced GAN into image fusion, the unsupervised learning capabilities of GAN have gained significant attention in this field. However, a single discriminator architecture often leads to modality weight imbalance, prompting subsequent studies to refine adversarial mechanisms for improved fusion control. For instance, GANMcC [21] proposes a multiclass constraint discriminator, which forces the fused result to retain a balanced distribution of both modalities.

In recent years, Transformers and hybrid architectures have become a research hotspot due to their superior global modeling capabilities. CDDFuse [22] introduces a correlation-driven feature decomposition network, incorporating Lite Transformer blocks to extract low-frequency global features, thereby achieving cross-modal feature disentanglement. EMMA [23] designs a Restormer-CNN hybrid block, integrating geometric priors from natural imaging into self-supervised training. These methods leverage multiscale feature collaboration and prior knowledge embedding, providing new perspectives for dynamic weight allocation and detail fidelity in image fusion.

B. Diffusion Models

A type of generative model called the diffusion model has just surfaced in the deep learning sector, and it currently shows notable benefits in image super-resolution, image generation, and image restoration. This probabilistic model perturbs the input data by incrementally adding Gaussian noise during forward diffusion before learning to reverse the process and retrieve the desired noise-free data from the noisy samples [24]. During this procedure, features of different spatial frequencies are extracted from the images. This feature opens up the possibility of applying diffusion models to the field of image fusion. Diffusion [25] pioneered the use of diffusion modeling for infrared-visible image fusion by extracting multichannel diffusion features through a denoising network and generating three-channel color fused images through a multichannel fusion module. The model introduces intensity loss and multichannel gradient loss to preserve intensity and texture information, hence improving the color accuracy of the fused images.

C. Attention Mechanisms

Attention mechanisms have revolutionized various fields [26], [27], [28] in deep learning by allowing models to concentrate on the most pertinent parts of the input data. Initially introduced in the context of machine translation [29], attention mechanisms have since been adapted to numerous other tasks, including image processing, where they help models selectively concentrate on important features. In image fusion, attention mechanisms play a crucial role by enhancing the extraction and fusion of features from multiple sources. Squeeze-and-excitation networks (SENets) [30] introduced channel attention mechanisms that adaptively recalibrate the channel-wise feature responses, significantly improving model performance on image classification tasks. Similarly, the convolutional block attention module (CBAM) [31] extended this concept by integrating both channel and spatial attention, further refining feature representation.

D. Invertible Neural Networks

Invertible Neural Networks have garnered significant attention in deep learning due to their unique capability of achieving a fully reversible mapping between input and output spaces. This characteristic enables the INN to retain complete information about the input data, thereby facilitating high-quality image reconstruction and synthesis. The concept of the INN was first introduced with the nonlinear independent components estimation (NICE) model [32]. The NICE model leveraged a series of coupling layers that ensured the invertibility of the network, allowing for efficient and lossless data transformations. The RealNVP model [33] enhanced the NICE architecture by incorporating a more sophisticated coupling layer design, which further optimized the network's performance in generative tasks. Ardizzone et al. [34] made significant contributions to the understanding and application of INNs in their work on normalized flow models. Radev et al. [35] showcased the efficacy of INNs in maintaining detailed and accurate image representations. INet [36] leverages INN to achieve lossless information processing in multimodal medical image fusion. In addition to image fusion and reconstruction, INNs have found applications in image hiding and other areas of image processing.

III. Proposed Method

We present the workflow of our DMNet and the specific structure of the individual modules in this section. Fig. 2 depicts the workflow in detail.

A. Overview

The proposed method is a hierarchical TST fusion framework. In Training Stage I, a dual-branch parallel training strategy is adopted, where visible and infrared images are separately fed into an encoder–decoder network and an independently optimized diffusion model. The encoder follows a three-layer progressive feature abstraction architecture: the improved DTB and DIB are designed to enhance long-range dependency modeling and nonlocal feature disentanglement, respectively; the CSAB module selectively

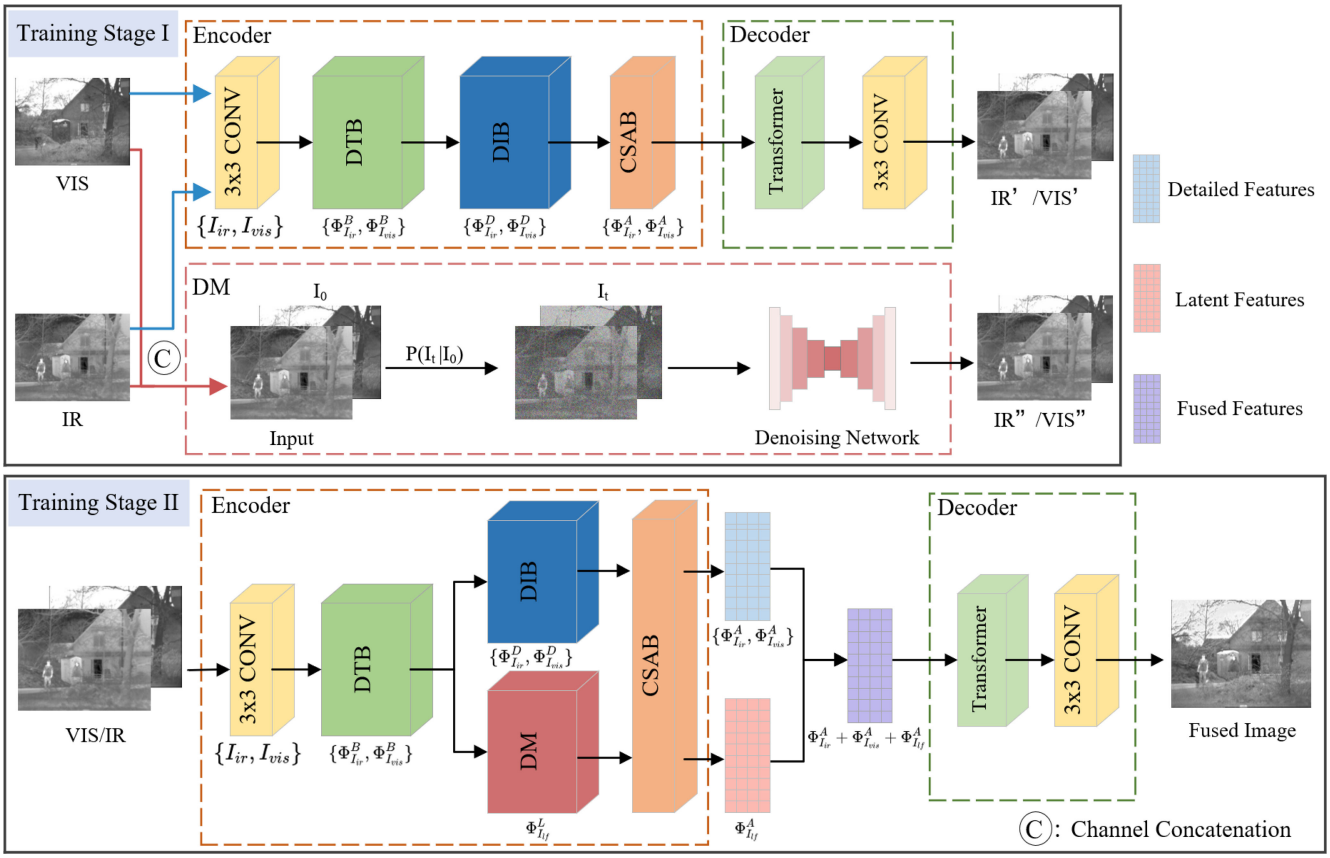

Fig. 2. Framework of our DMNet method. (a) AE structure and diffusion module for feature decomposition and reconstruction of the source images in the Training Stage I. (b) AE structure for obtaining the fused images in the Training Stage II.

focuses on important feature maps through a joint channel and spatial attention mechanism. The diffusion model reconstructs images through a noise-adding and denoising process while extracting latent representations via the denoising network. In Training Stage II, after passing through the DTB module, the input images are separately processed by DIB and the diffusion model (DM). DIB extracts detailed features, while DM captures cross-modal latent representations. The features from both pathways are then guided by CSAB before being fed into the decoder to generate the final fused image.

B. DTB and DIB

For clarity of presentation in formulation, we define some notation. The paired IR and VIS images of the aligned inputs are represented as $I_{\text{ir}} \in \mathbb{R}^{H\times W}$ and $I_{\text{vis}} \in \mathbb{R}^{H\times W\times 3}$.

The structure of the DTB and the DIB is shown in Fig. 3. We use the DTB to extract global basic features $\{\Phi^B_{I_{\text{ir}}}, \Phi^B_{I_{\text{vis}}}\}$ applying self-attention on the feature dimension of infrared and visible inputs from high-resolution input images $\{I_{\text{ir}}, I_{\text{vis}}\}$, i.e.,

$$\Phi^B_{I_{\text{ir}}} = \mathcal{B}(I_{\text{ir}}), \quad \Phi^B_{I_{\text{vis}}} = \mathcal{B}(I_{\text{vis}}) \tag{1}$$

where $\mathcal{B}(\cdot)$ and $\mathcal{D}(\cdot)$ represent the DTB and the DIB, respectively.

Given the texture and edge in the basic features are quite significant for the process of image fusion, we want to preserve as rich detail information as feasible after extracting the global features, thus we utilize the DIB with affine coupling layers [22]. The DIB enables the input and output features generate each other, so that the input features can be more effectively maintained. Therefore, it is regarded as a lossless feature extraction block and well suited for this task. The DIB then extracts the detailed information features from the global basic features with the formula

$$\Phi^D_{I_{\text{ir}}} = \mathcal{D}\left(\Phi^B_{I_{\text{ir}}}\right), \quad \Phi^D_{I_{\text{vis}}} = \mathcal{D}\left(\Phi^B_{I_{\text{vis}}}\right). \tag{2}$$

As shown in Fig. 3, the DTB contains five transformer layers and the DIB has three INN layers where each layer's output cascade is used as the input to the next layer. Each layer outputs the same dimension. And ω and ψ are the tuning parameters that we initially set to $1/5$ and $1/3$, respectively.

C. Denoising Network

We concatenate the source image in the channel dimension and use the two-channel image as the input to the diffusion module. The main core process of the diffusion module is shown in Fig. 2. Gaussian noise is gradually added to the image in the forward process, and finally presents a nearly pure noise state $\mathcal{P}\left(I_t \mid I_{t-1}\right)$. The denoising process $\mathcal{Q}\left(I_{t-1} \mid I_t\right)$ gradually predicts and eliminates the noise through the denoising network in the reverse process. Finally, the latent features of the input image are extracted by the denoising network.

Motivated by previous works such as Diffusion [25] and DANet [37], we adopt the channel-wise concatenation of

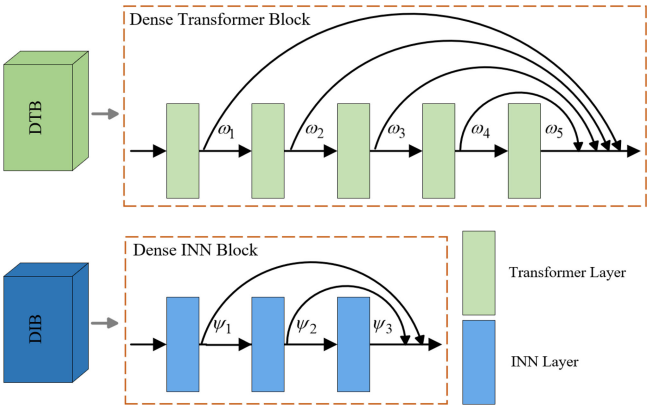

Fig. 3. Structures of the DTB and dense INN block.

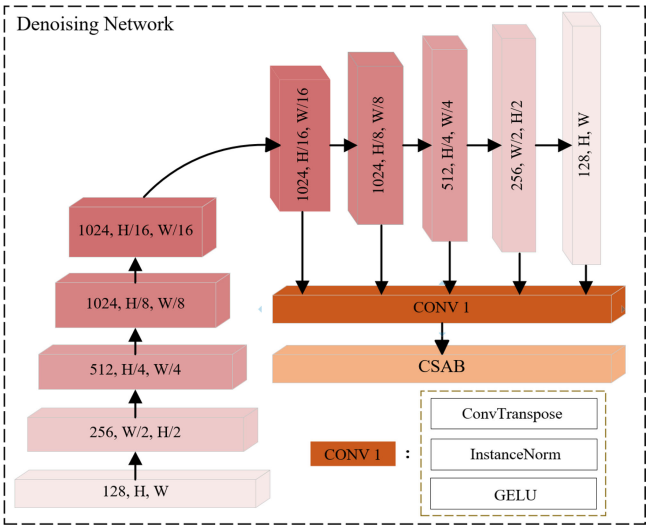

Fig. 4. Architecture of our proposed denoising network.

infrared and visible images as the input to the DDPM. This design aims to enable the diffusion process to jointly perceive both modalities. By fusing information at the channel level, the forward noise addition and reverse denoising processes can simultaneously capture modality-specific and complementary patterns. This strategy allows the model to encode cross-modal dependencies during the diffusion process and extract more representative and informative latent features.

Specifically, the two-channel image $I_0 \in \mathbb{R}^{H\times W\times 2}$, which is concatenated from the source images, utilizes the bidirectional process in the DDPM [25] to establish the distribution of the two-channel data. The forward process gradually adds noise over T time steps, while the reverse process removes noise from the added data to recover the original data gradually. Through both the forward and reverse processes, the diffusion module learns the shared latent structure of the source images and extracts features of different spatial frequencies [25].

Forward Diffusion Process: The forward process of the diffusion module can be regarded as a Markov chain [38] driven by nonequilibrium thermodynamics principles. Gaussian noise is progressively introduced to the data samples at each time step t (from 0 to T). The process of adding noise can be expressed as the recursive formula

$$\mathcal{P}(I_t \mid I_{t-1}) = \mathcal{N}\left(\sqrt{\alpha_t} I_{t-1}, 1-\alpha_t\right) \tag{3}$$

where N denotes the normal distribution. I_{t-1} and I_t denote the noisy two-channel images generated after adding Gaussian noise $t-1$ and t steps, respectively. The variance table α_t is used to determine the variance of the Gaussian noise injected at time step t. More specifically, after the first time step, I_1 can be expressed as

$$I_1 = \sqrt{1-\alpha_1} \times \epsilon_1 + \sqrt{\alpha_1} \times I_0 \tag{4}$$

where $\epsilon_t \in \mathbb{R}^{H\times W\times 2}$ is the Gaussian noise obeying a standard normal distribution at the moment of time step t. From (3) and (4) it can be deduced that

$$I_t = \sqrt{1-\bar{\alpha}_t} \times \epsilon + \sqrt{\alpha_1} \times I_0 \tag{5}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Based on the above deduction, we can define the diffusion process from I_0 to I_t as

$$\mathcal{P}(I_t \mid I_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t} I_0, 1-\bar{\alpha}_t\right). \tag{6}$$

Reverse Diffusion Process: The original two-channel image is obtained by applying a series of denoising operations using a neural network [39]. Given that the transition from time step t to $t-1$ is a stochastic process, we employ Bayes' theorem [35] to derive the image I_{t-1} of the previous time step from the known image I_t at the time step t. This process can be formulated as

$$Q(I_{t-1} \mid I_t) = \frac{P(I_t \mid I_{t-1}) P(I_{t-1} \mid I_0)}{P(I_t \mid I_0)} \tag{7}$$

based on (3), (5), and (6), it can be deduced that

$$\mathcal{Q}(I_{t-1} \mid I_t) = \mathcal{N}\left(\frac{\sqrt{\alpha_t}\beta_{t-1}}{\beta_t} I_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\beta_t} I_0, \frac{(1-\alpha_t)\beta_{t-1}}{\beta_t}\right) \tag{8}$$

where $I_0 = (I_t - \sqrt{\beta_t} \times \epsilon)/\sqrt{\bar{\alpha}_t}$, and $\beta_t = 1-\alpha_t$.

Structure of the Denoising Network: The purpose of the denoising network is to predict and eliminate the noise added during forward processing. The network structure of the denoising network we designed is shown in Fig. 4. It follows the U-Net architecture [40] used in SR3. In SR3 [25], the expansion path of the backbone consists of five convolutional layers that generate output feature maps at different scales. This multiresolution structure enhances the model's ability to reconstruct the fused output, resulting in rich latent representations with high structural fidelity and perceptual quality. We use one convolutional layer to fuse the multichannel diffusion features generated by the five stages of the denoising network and finally generate the latent feature maps $\Phi^L_{I_{\text{lf}}}$. The denoising network extracts the latent features from the global basic features with the formula

$$\Phi^L_{I_{\text{lf}}} = \mathcal{L}\left(\text{Conv}(\mathcal{C}(\Phi^B_{I_{\text{ir}}}, \Phi^B_{I_{\text{vis}}}))\right) \tag{9}$$

where $\mathcal{L}(\cdot)$ represents the DM, $\text{Conv}(\cdot)$ donates a convolution operation, $\mathcal{C}(\cdot)$ refers to channel concatenation. The input dimensions in the second stage are the same as in the first stage.

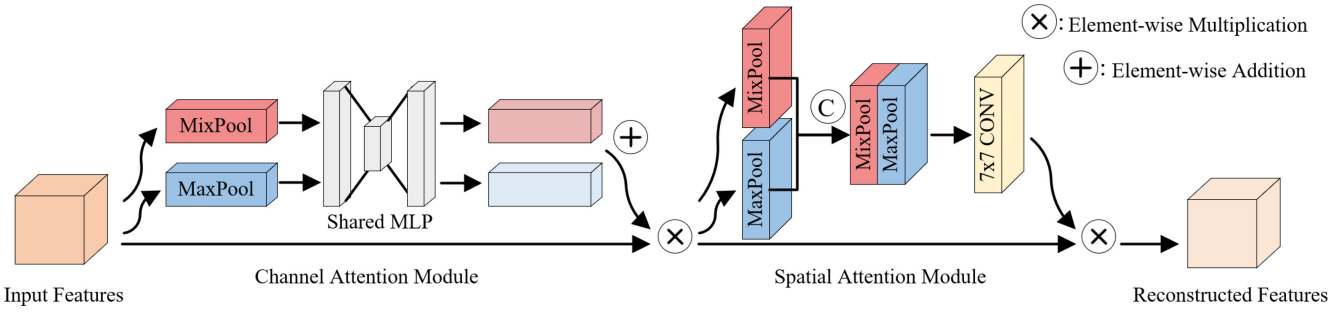

Fig. 5. Structures of the CSAB.

D. CSAB

Inspired by [31], we design the CSAB. The CSAB is a CNN-based attention module designed to enhance the representation capacity and performance of the model. By selectively attending to important feature maps and suppressing unimportant features through the joint channel and spatial attention mechanism, our CSAB improves the model's capacity to capture crucial data. As shown in Fig. 5, the Channel Attention Module and the Spatial Attention Module make up its two primary submodules.

The channel attention module generates two different description vectors by performing global mixed pooling and global max pooling on the input feature maps. They are then fused through a shared multilayer perceptron (MLP). Ultimately, the activation function sigmoid creates the channel attention map, which enhances the feature map's ability to discriminate.

Using global max pooling and global mixed pooling on the input feature maps in the channel dimension, the spatial attention module creates two distinct spatial attention maps. Then, they are fused by convolution operation. Ultimately, the activation function sigmoid creates the spatial attention maps, improving the model's capacity to extract key spatial information.

The CSAB is employed to improve the expressiveness of the features. The detailed features extracted by DIB and the latent feature maps extracted by DM are, respectively, feature-enhanced by the CSAB module, and the expressions are

$$\Phi^A_{I_{\text{ir}}} = \mathcal{A}\left(\Phi^D_{I_{\text{ir}}}\right),\ \Phi^A_{I_{\text{vis}}} = \mathcal{A}\left(\Phi^D_{I_{\text{vis}}}\right),\ \Phi^A_{I_{\text{lf}}} = \mathcal{A}\left(\Phi^L_{I_{\text{lf}}}\right) \tag{10}$$

where $\mathcal{A}(\cdot)$ represent the CSAB.

E. Encoder

The Encoder has five components: 1) a classical CNN block; 2) a DTB; 3) a DIB; 4) a CSAB; and 5) a DM. Each of these components is designed to progressively extract and refine features from the input image, enabling the network to capture both low-level and high-level representations essential for effective image fusion.

In Training Stage I: We embed the input images by a simple convolution, then the global feature information $\{\Phi^B_{I_{\text{ir}}}, \Phi^B_{I_{\text{vis}}}\}$ is extracted by the DTB, the detailed features $\{\Phi^D_{I_{\text{ir}}}, \Phi^D_{I_{\text{vis}}}\}$ are extracted after the DIB, and finally the reconstructed important features $\{\Phi^A_{I_{\text{ir}}}, \Phi^A_{I_{\text{vis}}}\}$ are obtained by the CSAB.

In Training Stage II: The global feature information $\{\Phi^B_{I_{\text{ir}}}, \Phi^B_{I_{\text{vis}}}\}$ extracted by the DTB is used to extract detail features $\{\Phi^D_{I_{\text{ir}}}, \Phi^D_{I_{\text{vis}}}\}$ and latent features $\Phi^L_{I_{\text{lf}}}$ through the DIB and DM, respectively. Finally, the reconstructed detail features $\{\Phi^A_{I_{\text{ir}}}, \Phi^A_{I_{\text{vis}}}\}$ and latent features $\Phi^A_{I_{\text{lf}}}$ are obtained through the CSAB block.

F. Decoder

The Decoder plays a crucial role in reconstructing the source and fused images from cross-modal features. Given the inherent differences in modality between infrared and visible images, it is essential for the Decoder to possess both strong global modeling capabilities and precise local detail reconstruction.

To achieve this, we design the Decoder as a hybrid structure that combines Transformer and CNN modules. The Transformer blocks enable the Decoder to capture global contextual information and long-range dependencies, while CNNs preserve local textures. This hybrid design has been successfully adopted by recent state-of-the-art methods such as CDDFuse [22] and DANet [37], validating its effectiveness in fusion tasks.

Since the inputs are cross-modal, we aim to ensure stable training of the model so that the Decoder may make use of the features that the Encoder generated [41] for accurate reconstruction [42]. To achieve this, we maintain a consistent Decoder structure across the two stages of training. This approach reduces the risk of information loss and consequently enhances the quality of the fused images.

In Training Stage I: The important detailed features of the decomposed infrared and visible image pairs are, respectively, used as inputs to the Decoders, and the reconstructed infrared and visible images are used as outputs, which is formulated as

$$I_{\text{ir}'} = \mathcal{DE}\left(\Phi^A_{I_{\text{ir}}}\right), I_{\text{vis}'} = \mathcal{DE}\left(\Phi^A_{I_{\text{vis}}}\right) \tag{11}$$

where $\mathcal{DE}(\cdot)$ represents the Decoder.

In Training Stage II: We add the important detailed features and the important latent features extracted by the CSAB as the input to the Decoder, and the fused image is used as the output, which is formulated as

$$F = \mathcal{DE}\left(\Phi^A_{I_{\text{ir}}} + \Phi^A_{I_{\text{vis}}} + \Phi^A_{I_{\text{lf}}}\right). \tag{12}$$

G. Two-Stage Training

State-of-the-art supervised learning methods often struggle with image fusion tasks due to the scarcity of ground truth. To overcome the challenge, we have drawn inspiration from [22] and employ a two-stage end-to-end learning approach to train the DMNet. This helps to effectively address the difficulties associated with image fusion. Specifically, the modules in the network are trained in the first stage to ensure that each module extracts and processes features efficiently so that better fusion results can be obtained in the second stage. This strategy effectively improves the robustness of the model and the quality of the fused images.

In Training Stage I: The sets of visible and infrared images $\{I_{\text{ir}}, I_{\text{vis}}\}$ are processed through the denoising network to obtain latent feature maps. Meanwhile, features $\{\Phi^A_{I_{\text{ir}}}, \Phi^A_{I_{\text{vis}}}\}$ are extracted from the infrared and visible image pairs $\{I_{\text{ir}}, I_{\text{vis}}\}$ that are fed into the Encoder. After that, the original infrared and visible images $\{I_{\text{ir}'}, I_{\text{vis}'}\}$ are, respectively, reconstruct using these features by the Decoder.

In Training Stage II: We retain the encoder and decoder parameters obtained from Stage I as the initial weights and continue to train the entire network jointly. Pairs of infrared-visible images $\{I_{\text{ir}}, I_{\text{vis}}\}$ are fed to an Encoder which is almost well-trained for extracting features across various frequency domains. The reconstructed detail features obtained are summed with the reconstructed latent features to form the fused features $\{\Phi^A_{I_{\text{ir}}} + \Phi^A_{I_{\text{vis}}} + \Phi^A_{I_{\text{lf}}}\}$. The final fused image F is generated by processing the fused features by the Decoder. Due to the two-stage learning process, our network enhances the quality of the fused image by more effectively capturing the finer details of the source images.

H. Loss Function

The loss function is not the same in the two training stages.

In Training Stage I: The loss of AE $\mathcal{L}_{au}$ is

$$\mathcal{L}_{au} = \mathcal{L}_{vi} + \gamma_1 \mathcal{L}_{\text{ir}} \tag{13}$$

where $\mathcal{L}_{vi}$ and $\mathcal{L}_{\text{ir}}$ denote the losses incurred during reconstruction for visible and infrared images [22], respectively. γ_1 is the tuning parameter. And the reconstruction losses are represented as

$$\mathcal{L}_{vi} = \mathcal{L}^I_{\text{int}}\left(I_{vi}, I_{vi'}\right) + \lambda \mathcal{L}_{\text{SSIM}}\left(I_{vi}, I_{vi'}\right) \tag{14}$$

$$\mathcal{L}^I_{\text{int}}\left(I_{vi}, I_{vi'}\right) = || \, I_{vi} - I_{vi'} \, ||_2^2 \tag{15}$$

$$\mathcal{L}_{\text{SSIM}}\left(I_{vi}, I_{vi'}\right) = 1 - \text{SSIM}\left(I_{vi}, I_{vi'}\right) \tag{16}$$

where $\text{SSIM}(\cdot)$ denotes the structural similarity index [43]. λ is the tuning parameter. We can get $\mathcal{L}_{\text{ir}}$ in the same way.

The loss of diffusion denoising process $\mathcal{L}_{dm}$ is defined as

$$\mathcal{L}_{dm} = || \, \xi - \eta\left(\sqrt{\bar{\alpha}_t} I_0 + \sqrt{1-\bar{\alpha}_t}\xi, t\right) \, ||_2 \tag{17}$$

where ξ represents sampling noise following a standard normal distribution, $\eta(\cdot, \cdot)$ denotes the denoising network. And the inputs of $\eta(\cdot, \cdot)$ are the noisy image I_t and the time step t.

In Training Stage II: The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{II}_{\text{int}} + \mathcal{L}_{dm} + \gamma_2 \mathcal{L}_{gd} \tag{18}$$

where $\mathcal{L}^{II}_{\text{int}} = (1/HW) \, || \, F - \max\left(I_{\text{ir}}, I_{\text{vis}}\right)||_1$, the intensity loss $\mathcal{L}^{II}_{\text{int}}$ helps to retain the salient intensity features. The definition of $\mathcal{L}_{dm}$ is the same as that in the first stage. γ_2 is the tuning parameter. $\mathcal{L}_{gd}$ is a gradient-based structural loss that preserves more prominent edges and contours in the fused image by aligning its gradient map with the stronger gradient components from the source images. $\mathcal{L}_{gd}$ is defined as

$$\mathcal{L}_{gd} = \frac{1}{HW} \, ||| \, \nabla F \, | - \max \, \left(| \, \nabla I_{\text{ir}} \, |, | \, \nabla I_{\text{vis}} \, |\right) ||_1 \tag{19}$$

where ∇ denotes the Sobel gradient operator.

IV. Experiments

In this section, we first introduce the experimental details, including dataset selection, training settings, benchmarks, and evaluation metrics. Subsequently, we conduct both quantitative and qualitative analyses on five publicly available infrared-visible datasets and three medical image datasets to evaluate the proposed model. To further validate the performance of our model and the superiority of its network design, we compare it with eight state-of-the-art models. Finally, comprehensive ablation experiments are conducted to demonstrate the rationality and effectiveness of the designed network and the proposed modules.

A. Datasets and Metrics

Datasets: Our experiments utilize infrared and visible image pairs from TNO [44], MSRS [45], RoadSence [46], and the LLVIP dataset [47] to validate our fusion model. The MSRS training set, which consists of 1083 pairs of visible and infrared images, is used to train the model we propose. Then, 25 pairs of TNO, 361 pairs of MSRS, 221 pairs of RoadScene and 3463 pairs of LLVIP test set images are used as the test dataset to validate the fusion performance.

Benchmarks: We compare our method with eleven state-of-the-art methods including DenseFuse [9], GANMcC [21], SDNet [48], U2Fusion [17], ReCoNet [16], MUFusion [19], CDDFuse [22], Diffusion [25], CrossFuse [18], EMMA [23], Diff-IF [49], DANet [37], and INet [36]. Among them, GANMcC is the fusion method based on generative models (GAN), SDNet, U2Fusion, ReCoNet, and INet use fusion methods based on CNN architectures, Diffusion is the fusion method based on diffusion models, while DenseFuse, MUFusion, CDDFuse, CrossFuse, and EMMA are fusion methods based on AE architecture.

Evaluation Metrics: The objective metrics EN [50], SF [51], AG [52], SD [53], CC [54], SCD [55], MI [56], VIF [57], Qabf [58], and MS-SSIM [59] are employed to analyze the performance of image fusion methods. EN, SF, AG, and SD are nonreference evaluation metrics, which do not need the fused image to calculate. CC, SCD, MI, VIF, Qabf, and MS-SSIM are reference evaluation metrics, which need the fused to calculate.

B. Implement Details

In the preprocessing stage, we randomly crop the visible and infrared images into 128 × 128 blocks. To improve

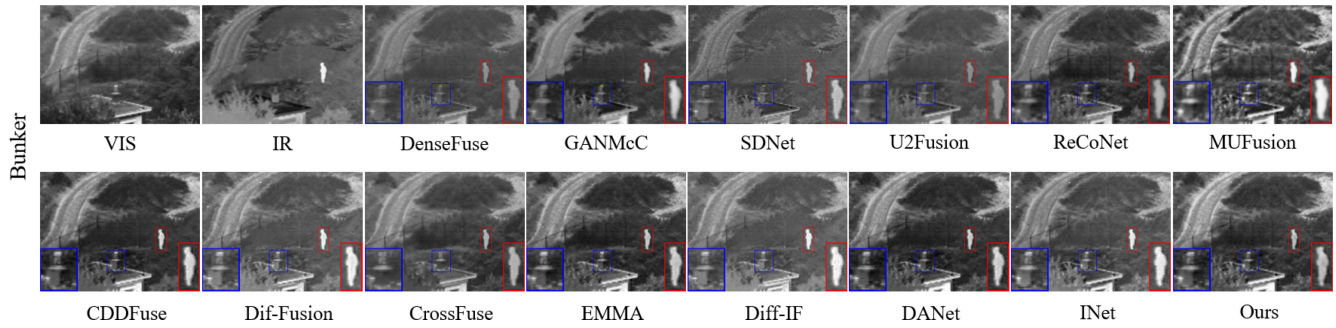

Fig. 6. Ten methods are compared qualitatively using the Bunker image pair from the TNO dataset.

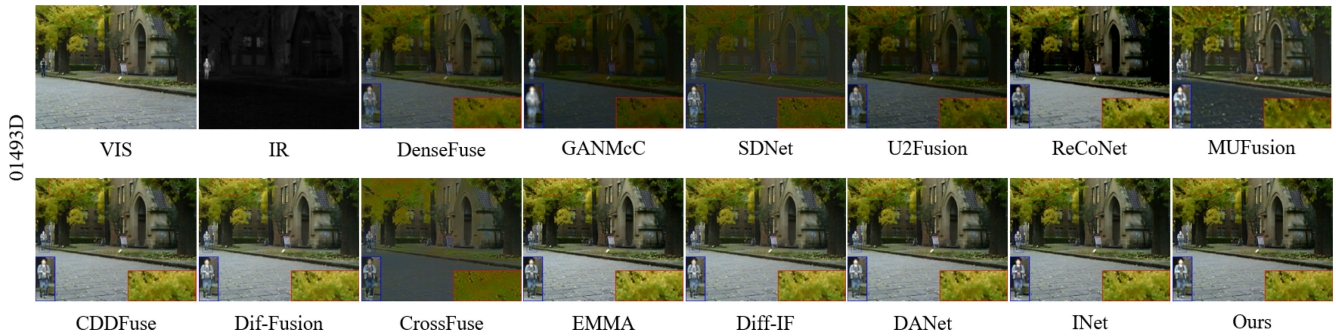

Fig. 7. Ten methods are compared qualitatively using the 01493D picture pair from the MSRS dataset.

the efficiency of training the diffusion module, multichannel diffusion features are generated by extracting the diffusion features generated at three time steps, namely 5, 50, and 100, as inspired by [25] and [39]. The first and second stages have 15 and 30 epochs, correspondingly, for a total of 45 epochs in AE training. To reduce the loss, we adopt the Adam optimizer with a 0.0001 initial learning rate. The suggested model is implemented using PyTorch. Every experiment carried out on an NVIDIA A100 Tensor Core GPU.

C. Fusion Performance Analysis

Figs. 6–9 give the visualization results for the four datasets, TNO, MSRS, RoadScene, and LLVIP, respectively. To get a clearer image of the performance of each method, we enlarge the target information and texture features in the fused images and visualize the results. Tables I–IV show the test results for each of the four datasets, where bold portions represent the best performance and underlined portions represent the second best performance. Overall, our model usually performs better than other methods.

Experiments on TNO Dataset: The qualitative and quantitative results on the TNO dataset are shown in Fig. 6 and Table I. Our method more effectively combines the thermal radiation information in the infrared image with the detailed texture information in the visible image, and the fused image has not only obvious infrared targets but also detailed texture information in the visible image. The important features of the two source images are not lost. A person in a bush is highlighted in the infrared picture in Fig. 6, and the soldier is retained in the fused image generated by all methods. The target person in DenseFuse, U2Fusion, and ReCoNet is darker. Some artifacts are observed along the edges of the soldier in the results produced by EMMA, Diff-IF, and INet. The edges of the turret in GANMcC, SDNet, CDDFuse, Diffusion, CrossFuse, and DANet are blurred, and the detail information is lost. The fusion image generated by MUFsion is somewhat sharpened. In addition, our network obtains optimal results in seven metrics and suboptimal results in two metrics. Although our experimental results are slightly lower than other methods on the MI metrics, this does not fully reflect the superiority of our method. Our method excels in visual quality and detail retention, which is verified in other critical metrics.

Experiments on MSRS Dataset: In Fig. 7, our method does better in the aspect of retaining more texture detail features from the source images. Specifically, in the red boxes of the fused images generated by DenseFuse, GANMcC, SDNet, ReCoNet, CrossFuse, and Diff-IF, the information of the leaves is lost or the texture information of the leaves becomes blurred. The fused images generated by GANMcC, U2Fusion, SDNet, ReCoNet, and CrossFuse are too dark and the detail information of the visible image is lost. The fused image generated by MUFsion is sharpened severely and does not match the human eye perception. The color of the human figures in the fused images generated by DANet and INet is distorted. Only the fused images generated by CDDFuse and EMMA closely resemble our results. Table II shows that our method obtains optimal results on seven metrics and suboptimal results on the remaining three metrics. Our method not only enhances the retention of critical details, but also significantly improves the overall visual quality of the fused images.

Experiments on RoadScene Dataset: We utilize the Roadscene dataset consisting of 221 image pairs of IR and VIS to verify the effectiveness of our method. The fused

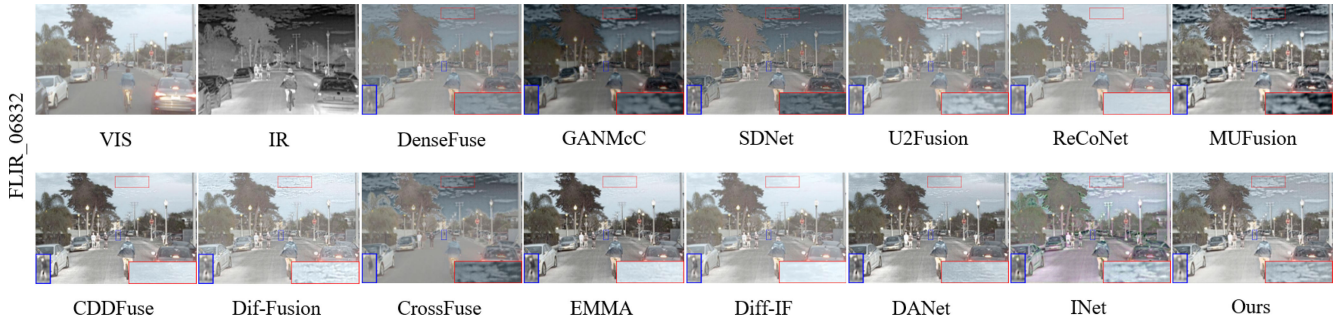

Fig. 8. Qualitative comparison of ten methods on the FLIR_06832 image pair from the RoadScene dataset.

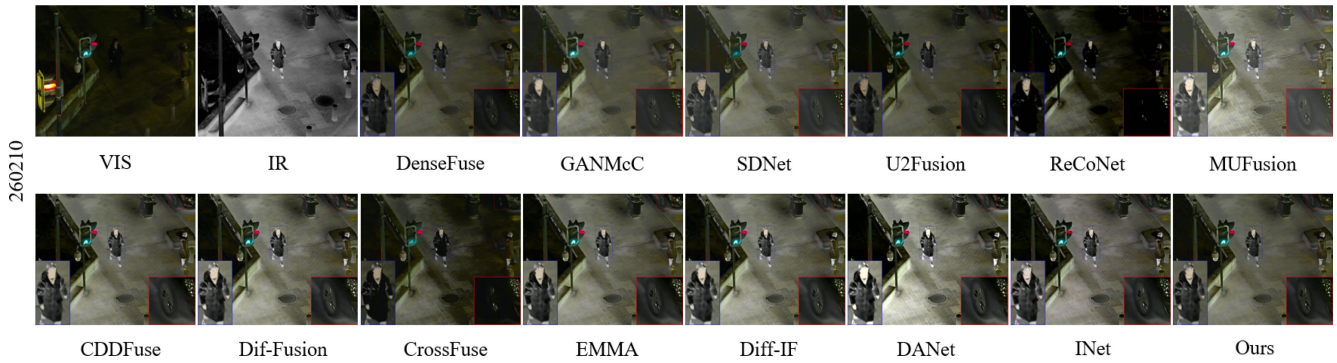

Fig. 9. Qualitative comparison of ten methods on the 260210 image pair from the LLVIP dataset.

TABLE I
EXPERIMENTAL RESULTS ON THE TNO DATASET [44]

Methods	EN	SF	AG	SD	CC	SCD	MI	VIF	Qabf	MS-SSIM
DenseFuse [9]	6.327	6.906	2.635	25.108	0.496	1.568	2.351	0.572	0.361	0.999
GANMcC [21]	6.646	6.343	2.551	32.225	0.492	1.657	2.342	0.512	0.281	0.967
SDNet [48]	6.333	10.09	3.831	26.732	0.471	1.422	2.268	0.539	0.438	0.953
U2Fusion [17]	6.482	6.511	2.699	32.608	0.497	1.631	2.319	0.524	0.356	0.993
ReCoNet [16]	6.668	7.744	3.289	40.519	0.469	1.711	2.436	0.526	0.368	0.972
MUFusion [19]	7.092	10.311	**4.807**	44.907	0.443	1.539	1.963	0.521	0.352	0.829
CDDFuse [22]	7.096	11.165	4.594	44.927	0.475	1.648	3.332	**0.681**	0.471	0.997
Dif-Fusion [25]	6.942	11.034	4.446	40.349	0.457	1.602	2.501	0.592	0.469	0.946
CrossFuse [18]	6.602	10.139	3.648	31.539	0.426	1.413	**4.264**	0.579	0.473	0.906
EMMA [23]	7.094	10.983	4.587	44.839	0.459	1.658	3.126	0.583	0.472	0.961
Diff-IF [49]	6.933	10.898	4.388	40.827	0.452	1.573	3.659	0.591	0.477	0.942
DANet [37]	7.059	10.968	4.130	45.016	0.486	1.807	2.227	0.546	0.469	1.001
INet [36]	6.985	10.214	4.126	39.621	0.483	1.702	2.402	0.588	0.479	0.992
Ours	**7.129**	**11.199**	4.684	**45.134**	**0.504**	**1.859**	2.511	0.593	**0.483**	**1.012**

results of a quintessentially pair of IR and VIS images are shown in Fig. 8. The fused images from ReCoNet, CDDFuse, Diffusion, EMMA, Diff-IF, and DANet are too high contrast, resulting in the loss of texture information of the cloud in the red boxes. The fused images from by DenseFuse, GANMcC, SDNet, and CrossFuse are too dark, making the character information in the blue boxes unclear. The fused images generated by MUFsion are still badly sharpened. The color of the fused image generated by INet is distorted. The fused image generated by U2Fusion retains the texture detail information to some extent, but the target is not clear and the boundary is blurred. In contrast, our method not only pays good attention to the luminance information of the source images, but also preserves the rich detailed texture information. Table III shows that our method obtains optimal results on eight metrics and suboptimal results on one metric. Our method is comparatively lower than some other methods for the metric MI. This is because the feature fusion strategy of our method focuses on preserving luminance information and texture details. Due to the large difference in luminance information between the IR and VIS source images on the RoadScene dataset, the fused features can be considered as a tradeoff of the original feature maps, which results in lower mutual information between our fused image and the infrared image. However, our method achieves better performance than other methods from the perspective of multimetric evaluation.

Experiments on LLVIP Dataset: LLVIP is a IR-VIS paired dataset for low-light vision [47]. We validate the performance of all fusion method on the test set consisting of 3463 IR-VIS image pair. In Fig. 9, the fused images generated

TABLE II
EXPERIMENTAL RESULTS ON THE MSRS DATASET [45]

Methods	EN	SF	AG	SD	CC	SCD	MI	VIF	Qabf	MS-SSIM
DenseFuse [9]	6.518	7.528	2.683	28.172	0.505	1.272	3.151	0.655	0.378	0.921
GANMcC [21]	6.299	6.291	2.281	26.699	0.518	1.271	2.803	0.517	0.225	0.824
SDNet [48]	5.501	9.597	3.089	16.954	0.515	1.023	1.555	0.415	0.335	0.789
U2Fusion [17]	6.509	8.202	2.979	34.253	0.514	1.446	2.648	0.582	0.395	0.924
ReCoNet [16]	6.826	13.178	4.932	**57.002**	0.519	1.401	3.582	0.663	0.568	0.905
MUFusion [19]	6.646	11.098	4.389	33.688	0.544	1.147	1.561	0.527	0.434	0.859
CDDFuse [22]	7.096	13.791	5.063	53.849	0.518	1.478	4.309	**1.008**	0.651	0.987
Dif-Fusion [25]	7.419	13.976	4.815	49.878	**0.664**	1.457	3.422	0.721	0.555	0.949
CrossFuse [18]	6.282	10.638	3.541	24.131	0.496	0.919	3.436	0.538	0.451	0.777
EMMA [23]	7.375	14.227	4.516	54.139	0.517	1.445	4.963	0.912	0.708	0.992
Diff-IF [49]	7.365	14.203	5.010	53.525	0.524	1.503	5.078	0.913	0.696	0.990
DANet [37]	7.418	14.539	5.069	54.261	0.525	1.598	5.024	0.929	0704	0.994
INet [36]	7.386	13.438	4.517	51.981	0.539	1.558	3.669	0.905	0.703	1.001
Ours	**7.424**	**14.655**	**5.076**	54.455	0.546	**1.796**	**5.161**	0.934	**0.712**	**1.003**

TABLE III
EXPERIMENTAL RESULTS ON THE ROADSCENE DATASET [46]

Methods	EN	SF	AG	SD	CC	SCD	MI	VIF	Qabf	MS-SSIM
DenseFuse [9]	6.821	8.626	3.391	32.182	0.606	1.364	2.959	0.546	0.393	1.058
GANMcC [21]	7.237	9.145	3.836	43.882	0.644	1.613	2.857	0.513	0.359	1.053
SDNet [48]	7.052	12.871	5.177	37.594	0.578	1.192	3.039	0.517	0.335	1.065
U2Fusion [17]	6.998	8.945	3.709	37.222	0.644	1.499	2.847	0.552	0.397	1.059
ReCoNet [16]	7.052	9.129	3.825	41.391	0.621	1.539	3.106	0.543	0.383	1.057
MUFusion [19]	7.419	13.483	6.027	51.531	0.593	1.513	2.233	0.493	0.363	0.968
CDDFuse [22]	7.226	14.131	5.859	56.517	0.627	1.709	3.099	**0.621**	0.422	1.066
Dif-Fusion [25]	7.169	15.302	5.633	42.537	0.581	1.325	2.909	0.545	0.452	0.993
CrossFuse [18]	6.794	9.903	3.556	32.407	0.553	1.033	**4.686**	0.509	0.423	0.981
EMMA [23]	7.102	14.199	5.254	56.112	0.615	1.635	3.204	0.592	0.445	1.055
Diff-IF [49]	7.103	15.323	5.487	43.789	0.574	1.303	3.716	0.570	0.419	0.993
DANet [37]	7.403	15.465	6.006	56.106	0.643	1.812	2.912	0.590	0.451	1.065
INet [36]	7.315	15.654	5.439	44.711	0.603	1.434	2.787	0.585	0.402	1.051
Ours	**7.426**	**15.735**	**6.034**	**56.999**	**0.645**	**1.834**	2.981	0.595	**0.462**	**1.073**

TABLE IV
EXPERIMENTAL RESULTS ON THE LLVIP DATASET [47]

Methods	EN	SF	AG	SD	CC	SCD	MI	VIF	Qabf	MS-SSIM
DenseFuse [9]	6.832	8.428	2.511	33.993	0.623	1.235	2.904	0.739	0.394	0.985
GANMcC [21]	6.737	7.156	2.195	33.556	0.618	1.171	2.781	0.589	0.259	0.889
SDNet [48]	6.777	12.502	3.655	33.539	0.556	0.935	3.128	0.639	0.539	0.897
U2Fusion [17]	6.783	7.94	2.492	36.241	0.608	1.273	2.963	0.662	0.366	0.955
ReCoNet [16]	5.668	10.524	3.178	46.091	0.687	1.454	2.179	0.556	0.404	0.791
MUFusion [19]	7.002	10.211	3.569	39.615	0.638	1.043	2.532	0.679	0.458	0.901
CDDFuse [22]	7.133	12.248	4.007	44.949	0.685	1.583	**4.598**	**0.879**	0.536	1.007
Dif-Fusion [25]	7.159	12.302	3.933	42.537	0.581	1.325	2.909	0.545	0.512	0.993
CrossFuse [18]	6.674	11.604	3.703	37.634	0.661	1.113	4.352	0.698	0.579	0.915
EMMA [23]	7.069	11.911	3.943	46.046	0.684	1.568	3.669	0.753	0.574	1.017
Diff-IF [49]	7.123	12.457	3.791	45.888	0.666	1.495	4.197	0.761	0.539	0.978
DANet [37]	7.094	12.477	4.059	46.088	0.679	1.640	2.761	0.762	0.577	1.013
INet [36]	7.167	12.076	4.064	45.684	0.667	1.451	2.789	0.747	0.518	1.016
Ours	**7.261**	**12.868**	**4.086**	**48.665**	**0.718**	**1.703**	2.878	0.775	**0.592**	**1.031**

by DenseFuse, ReCoNet, and CrossFuse are darker and the texture detail information in the images are blurred. The fused images from MUFsion, CDDFuse, Diffusion, Diff-IF, and INet have lower contrast, which may lead to difficulties for subsequent tasks (e.g., object detection and image segmentation). We can observe in the red boxes that only our method preserves the car wheel texture and detail information well while retaining the appropriate luminance information. Table IV shows that our method obtains optimal results on eight metrics and suboptimal results on one metric. Compared with the suboptimal method, our SF and SD sores are improved by 0.366 and 2.574, respectively.

D. Ablation Studies

To validate the performance of our method, we conducted a series of ablation experiments to explore the effects of different modules, respectively. To verify the performance of TST, we only conduct training stage 2 in AE1. Meanwhile, to ensure fairness, we set the epoch of the training stage 2 to 45. In the ablation experiments, we adopt two strategies for the different modules: some modules are directly removed, while other complex modules (e.g., the diffusion module) are replaced by simple convolutional layers. We not only conduct separate ablation experiments for each module, but also combine ablation experiments for different modules. From

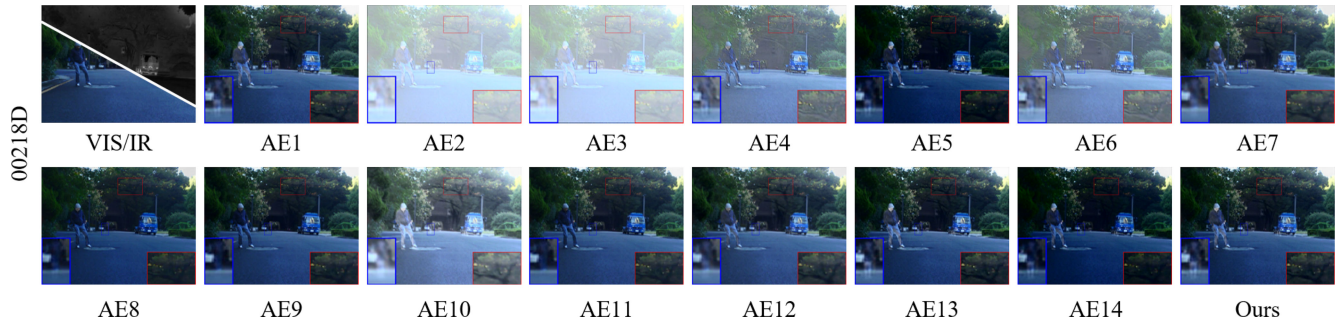

Fig. 10. Qualitative comparison of the ablation experiment for the fused results on the 00218D image pairs from the MSRS dataset.

TABLE V
ABLATION EXPERIMENT RESULTS IN THE DATASET OF MSRS. BOLD INDICATES THE BEST VALUE

	Baseline	TST	DTB	DIB	DM	CSAB	EN	SF	AG	SD	CC	SCD	MI	VIF	Qabf	MS-SSIM
AE1	✓	✗	✓	✓	✓	✓	7.321	13.728	4.951	51.444	0.527	1.519	4.406	0.841	0.671	0.992
AE2	✓	✓	✗	✗	✗	✗	6.298	10.229	4.155	48.989	0.524	1.298	2.989	0.522	0.348	0.900
AE3	✓	✓	✗	✗	✗	✓	6.429	10.958	4.174	50.291	0.529	1.498	3.736	0.713	0.594	0.923
AE4	✓	✓	✓	✗	✗	✗	6.408	8.334	2.775	25.793	0.435	0.791	3.355	0.400	0.208	0.647
AE5	✓	✓	✓	✓	✗	✗	7.312	14.543	5.003	52.768	0.536	1.735	4.125	0.801	0.646	1.004
AE6	✓	✓	✓	✗	✗	✓	6.978	11.005	4.144	39.464	0.439	0.794	3.283	0.501	0.434	0.821
AE7	✓	✓	✓	✗	✓	✗	6.351	8.438	2.855	26.633	0.437	0.803	3.339	0.404	0.216	0.657
AE8	✓	✓	✗	✓	✓	✗	7.063	11.693	4.066	45.143	0.515	1.278	4.573	0.768	0.653	0.958
AE9	✓	✓	✗	✓	✗	✓	7.334	13.463	4.623	52.045	0.496	1.449	4.139	0.914	0.701	0.990
AE10	✓	✓	✗	✗	✓	✓	7.209	11.074	4.202	51.146	0.527	1.488	3.701	0.711	0.593	0.995
AE11	✓	✓	✗	✓	✓	✓	7.104	13.447	4.643	50.299	0.486	1.055	**5.299**	0.929	0.706	0.989
AE12	✓	✓	✓	✗	✓	✓	6.972	11.724	4.561	37.488	0.424	0.743	2.712	0.411	0.409	0.771
AE13	✓	✓	✓	✓	✗	✓	7.406	14.459	5.029	53.428	0.533	1.637	3.999	0.825	0.655	**1.008**
AE14	✓	✓	✓	✓	✓	✗	7.243	13.995	4.929	53.341	0.544	1.692	3.941	0.755	0.622	0.987
Ours	✓	✓	✓	✓	✓	✓	**7.424**	**14.655**	**5.076**	**54.455**	**0.546**	**1.796**	5.161	**0.934**	**0.712**	1.003

TABLE VI
TEST RESULTS ON THE LLVIP DATASET USING DIFFERENT GRADIENT OPERATORS DURING TRAINING

Operators	EN	SF	AG	SD	VIF	Qabf	MS-SSIM
Canny	7.174	12.439	3.965	**48.703**	0.752	0.589	1.016
LoG	7.203	**12.891**	3.887	48.445	0.763	**0.602**	1.022
Roberts	6.981	12.176	3.766	47.924	0.651	0.549	1.001
Sobel	**7.261**	12.868	**4.086**	48.665	**0.775**	0.592	**1.031**

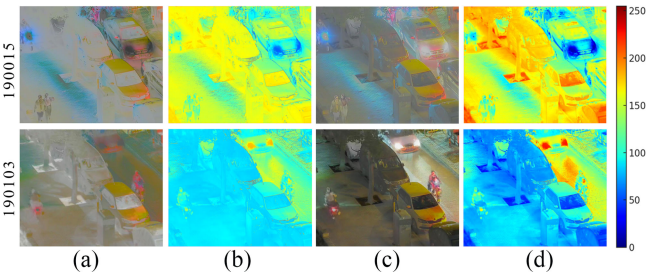

Fig. 11. Visualization of feature maps before and after CSAB module on image pairs 190015 and 190103 of the LLVIP dataset. The color bar represents the activation intensity. (a) $\Phi^D_{I_{\text{ir}}} + \Phi^D_{I_{\text{vis}}} + \Phi^L_{I_{\text{lf}}}$. (b) $HM(\Phi^D_{I_{\text{ir}}} + \Phi^D_{I_{\text{vis}}} + \Phi^L_{I_{\text{lf}}})$. (c) $\Phi^A_{I_{\text{ir}}} + \Phi^A_{I_{\text{vis}}} + \Phi^A_{I_{\text{lf}}}$. (d) $HM(\Phi^A_{I_{\text{ir}}} + \Phi^A_{I_{\text{vis}}} + \Phi^A_{I_{\text{lf}}})$.

AE1 to AE14 denote the ablation experiments of different modules, as detailed in Table V.

Qualitative Comparison: In Fig. 10, we can clearly observe that the fused images without the DIB module have low contrast and edge information is lost (e.g., AE7 and AE12). The texture detail information of the fused images without the DTB module is lost, e.g., the feature information is not visible in the blue boxes of AE8, AE9, and AE11. From AE13 and AE14, it can be seen that the lack of DM module and CSAB module leads to higher and lower brightness of the fused image, respectively. When both DIB module and DTB module are not simultaneously used, the quality of the fused image becomes very worse. The image contrast becomes low, brightness becomes high and a lot of texture information is lost as shown in AE2, AE3, and AE10. The absence of other combinations of different modules also degrades the quality of the fused image to varying degrees. By combining the advantages of all modules, our method well preserves the detailed texture and luminance hierarchy of the source image and presents better visual effects.

Quantitative Comparison: Synthesizing the ablation results in Table V, the efficacy and rationality of every module inside the network are validated by the network structure we designed, which yields the best average results in eight quality indicators and the second best average results in the remaining two. Although the MI metrics are slightly improved by removing the DTB module, in general, all other metrics are decreased. The MS-SSIM metric of AE13 with the DM module removed is only 0.005 higher than ours, yet all other metrics decrease. In summary, our DMNet has the best fusion performance.

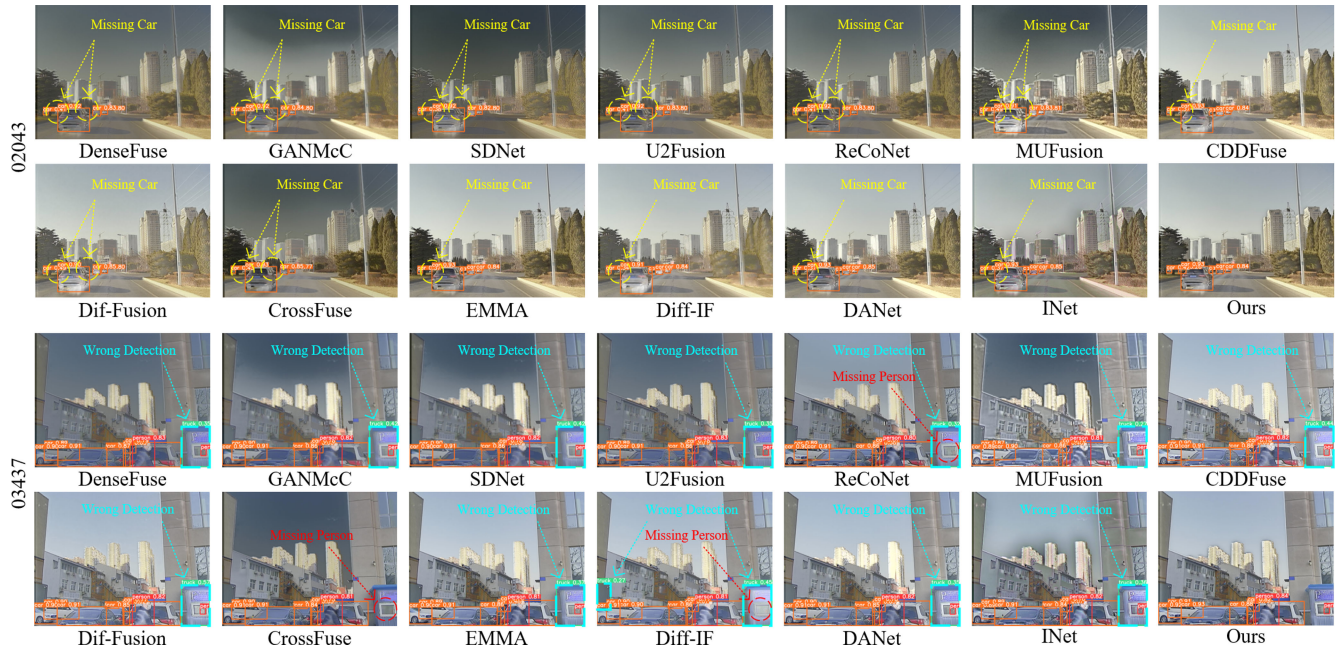

Fig. 12. Qualitative comparison of the object detection experiment for the fused results on the 02043 and 03437 image pairs from the M3FD dataset.

Operator Selection for Loss Function: We conducted ablation studies to compare the performance of the Sobel operator in the gradient loss function with other edge detection methods, such as Canny, LoG, and Roberts, in the image fusion training process. As shown in Table VI, the LoG operator achieves the best results in the SF and Qabf metrics, while the Canny operator performs best in the SD metric. In contrast, the Sobel operator consistently achieves the best or second-best performance across all metrics. Overall, the Sobel operator demonstrates the best comprehensive performance.

Visualization of Feature Maps: The results of the feature map visualization before and after the two pairs of CSAB modules are presented in Fig. 11, and their corresponding heat maps (HMs) are shown. In these HMs, we can observe that the CSAB module is able to enhance the focus on key feature regions. In the feature maps processed by the CSAB module, the higher weighted regions are usually focused on the salient structural and target information in the image, which are crucial for the image fusion task. For example, for the pedestrian and vehicle information in Fig. 11, the HM weights of these target regions are significantly increased after CSAB, indicating that the CSAB module is strengthening its focus on these critical information, thus improving the quality of the fused image. Compared to the feature maps not processed by the CSAB module, the HMs after CSAB show higher attention, especially in the edge and detail parts of the image, indicating that the module is able to effectively extract and retain fine-grained information in the image, which is crucial for further image analysis and understanding.

E. Experiments on Infrared-Visible Object Detection

We carry out IVIF object detection on the M3FD dataset, categorized into six groups, such as People, Car, Bus, Motor, Truck, and Lamp. The detector used for the object detection is the pretrained YOLOv5x [60] and the metric used for assessment is mAP@0.5.

Qualitative Comparison: The comparative analysis presented in Fig. 12 visually underscores the superiority of our method in object detection relative to other state-of-the-art approaches. Obviously, our competitors often miss at least one label or misidentify detection targets across various scenarios, failing to correctly detect all relevant objects. In contrast, our method is able to capture each target more comprehensively. In the figure 02043, other methods consistently fail to detect one or two occluded cars, whereas only our method successfully detects all of them. In the figure 03437, other methods either misclassify the security booth as a truck or fail to recognize the person sitting inside, while our method accurately detects these hard-to-recognize targets. This clear advantage highlights the effectiveness of our method in complex detection environments.

Quantitative Comparison: Our method gets the best performance in five categories of recognition, as indicated in Table VII, demonstrating its exceptional capabilities for object detection applications. Specifically, our method shows excellent performance in detecting the categories of person, car, and bus, obtaining the best detection results. This indicates that our method effectively captures valuable information in image fusion and reflects it in the fused image, thus greatly improving the accuracy of detecting hard-to-recognize targets.

F. Computational Complexity Discussion

We calculate the model parameters and average test time on the MSRS dataset for different methods, as shown in Table VIII. GANMcC processes stacked multisource images on CPU, leading to a long test time. Diffusion and DANet, both diffusion-based, also suffer from high time and parameter costs due to iterative denoising and deep architectures. In

TABLE VII
AP@0.5(%) VALUES FOR OBJECT DETECTION ON M3FD DATASET CONSISTING OF 4200 PAIRS OF INFRARED-VISIBLE IMAGES

Methods	People	Car	Bus	Motor	Trunk	Lamp	mAP
IR	49.90	59.83	28.48	10.30	9.34	4.15	27.00
VI	42.89	70.88	57.34	26.10	26.10	23.25	42.36
Den [9]	50.14	72.46	54.51	20.51	29.25	14.46	40.22
GAN [21]	46.75	70.75	53.07	16.83	27.76	11.34	37.75
SDN [48]	50.03	71.06	49.36	18.59	25.91	7.01	36.99
U2F [17]	49.27	71.36	55.61	20.09	14.13	**28.66**	39.85
ReC [16]	46.25	70.97	52.69	21.13	26.97	12.37	38.40
MUF [19]	45.84	70.56	55.59	20.78	27.67	11.65	38.68
CDD [22]	50.78	73.68	53.04	28.31	31.50	17.65	42.49
Dif [25]	50.97	73.18	54.24	29.95	27.92	17.09	41.89
Cro [18]	47.66	73.51	53.49	29.53	33.52	15.89	42.27
EMM [23]	48.84	72.43	54.46	25.45	29.43	13.45	40.68
DIF [49]	50.72	73.27	48.60	25.09	28.35	14.29	40.05
DAN [37]	51.27	74.48	58.73	29.87	32.92	18.06	44.22
INe [36]	51.78	73.55	57.31	24.29	25.51	19.49	41.98
Ours	**51.79**	**74.56**	**58.74**	**30.53**	**35.73**	18.67	**45.00**

TABLE VIII
PARAMETERS, AVERAGE TEST TIME, AND SF OF DIFFERENT METHODS ON THE MSRS DATASET

Methods	Framework	Params(M)	Test Time(s)	SF
Den [9]	CNN	0.297	0.319	7.528
GAN [21]	GAN	1.864	5.924	6.291
SDN [48]	CNN	0.067	0.498	9.597
U2F [17]	CNN	0.659	0.451	8.202
ReC [16]	CNN	**0.008**	0.268	13.178
MUF [19]	CNN	0.555	0.617	11.098
CDD [22]	CNN & Transformer	1.188	0.577	13.791
Cro [18]	CNN & Transformer	1.161	0.304	10.638
Dif [25]	CNN & DM	416.469	5.988	13.976
EMM [23]	CNN & Transformer	1.518	0.225	14.227
DIF [49]	DM	23.736	0.667	14.203
DAN [37]	Transformer & DM	391.046	2.018	14.539
INe [36]	INN	0.749	**0.215**	13.438
Ours	CNN & Transformer & DM	386.358	1.466	**14.655**

contrast, DIF achieves a smaller parameter count (23.736M) by using a shallower U-Net backbone, but its fusion quality is relatively worse. Our method balances accuracy and efficiency, it leverages a diffusion module only for latent feature extraction while excluding it from the final fusion step, significantly reducing complexity. This design balances model complexity and performance, as heavier models generally perform better on complex fusion tasks. Although diffusion-based methods tend to have larger parameter sizes, they generally achieve higher SF scores compared to other frameworks, indicating stronger capabilities in preserving detailed and structural information. Our method follows this trend and achieves the highest SF score (14.655) among all compared methods. Additionally, we optimize the number of sampling steps to cut computation time. Although our model contains 386M parameters, it is about 30M fewer than Diffusion and requires only one-quarter of its inference time. Our approach processes a pair of images in 1.5 s on average, achieving strong fusion performance while maintaining practical efficiency. The peak GPU memory consumption of our model during inference is approximately 9.94 GB, making it feasible for deployment on high-performance edge devices such as the NVIDIA Jetson AGX Orin (32 GB RAM). In future work, we will also consider introducing techniques such as model pruning and quantization to further extend the applicability of our model to lightweight embedded systems.

V. CONCLUSION

In this study, we propose a new AE architecture-based, unsupervised end-to-end infrared-visible image fusion method, called DMNet. Motivated by theoretical insights, we innovatively integrate the diffusion model, transformer, and invertible neural network into a unified and coordinated architecture, thereby achieving more comprehensive feature extraction. Specifically, the DTB module captures global-based information, the DM module extracts potential features, and the invertible neural network module is used to extract detailed features, with the CSAB module selectively focusing on important features. These extracted features are then fused together. Numerous experiments show that the fusion performance of our DMNet outperforms current state-of-the-art methods, and both qualitative and quantitative results validate the robustness, generality, and effectiveness of our DMNet. Since our network uses both the diffusion model and transformers, the computational volume is relatively large, and the demand for computational resources is relatively high. In the future, we will explore the improvement of our model to reduce the amount of required computational resources while maintaining the quality of the fused images.

REFERENCES

[1] D. Li, H. Zhang, N. Liu, and G. Wang, "Multiscale residual and attention guidance for low-light image enhancement in visual SLAM," *IEEE Internet Things J.*, vol. 11, no. 23, pp. 38370–38379, Dec. 2024.

[2] M. Yuan, X. Shi, N. Wang, Y. Wang, and X. Wei, "Improving RGB-infrared object detection with cascade alignment-guided transformer," *Inf. Fusion*, vol. 105, May 2024, Art. no. 102246.

[3] S. Liang, J. Lu, K. Zhang, and X. Chen, "Multiscale transformer hierarchically embedded CNN hybrid network for visible-infrared person reidentification," *IEEE Internet Things J.*, vol. 12, no. 7, pp. 9004–9018, Apr. 2025.

[4] H. Wei, X. Fu, Z. Wang, and J. Zhao, "Infrared/visible light fire image fusion method based on generative adversarial network of wavelet-guided pooling vision transformer," *Forests*, vol. 15, no. 6, p. 976, 2024.

[5] Z. Feng, X. Zhang, B. Zhou, M. Ren, and X. Chen, "NGST-net: A N-gram based Swin transformer network for improving multispectral and hyperspectral image fusion," *Int. J. Digit. Earth*, vol. 17, no. 1, 2024, Art. no. 2359574.

[6] M. González-Audícana, J. L. Saleta, R. G. Catalán, and R. García, "Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1291–1299, Jun. 2004.

[7] Z. Yuan and C. Shi, "MGN-Net: Multigranularity graph fusion network in multimodal for scene text spotting," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 25088–25098, Jul. 2024.

[8] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[9] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, pp. 2614–2623, 2019.

[10] O. Habash, S. Singh, R. Mizouni, and H. Otrok, "Multiple source localization in IoT: A conditional GAN and image-processing-based framework," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 7059–7070, Feb. 2024.

[11] Z. Zhang, D. Zhou, G. Sun, Y. Hu, and R. Deng, "DFTI: Dual-branch fusion network based on transformer and inception for space noncooperative objects," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, May 2024.

[12] C. Shi, L. Fang, H. Wu, X. Xian, Y. Shi, and L. Lin, "NiteDR: Nighttime image de-raining with cross-view sensor cooperative learning for dynamic driving scenes," *IEEE Trans. Multimedia*, vol. 26, pp. 9203–9215, Apr. 2024.

[13] X. Xiong et al., "Adaptive feature fusion and improved attention mechanism-based small object detection for UAV target tracking," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 21239–21249, Jun. 2024.
[14] H. Hu et al., "Video surveillance on mobile edge networks—A reinforcement-learning-based approach," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 4746–4760, Jun. 2020.
[15] Y. Zhong, F. Fei, and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, no. 2, 2016, Art. no. 25006.
[16] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, "ReCoNet: Recurrent correction network for fast and efficient multi-modality image fusion," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 539–555.
[17] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
[18] H. Li and X.-J. Wu, "CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach," *Inf. Fusion*, vol. 103, Mar. 2024, Art. no. 102147.
[19] C. Cheng, T. Xu, and X.-J. Wu, "MUFusion: A general unsupervised image fusion network based on memory unit," *Inf. Fusion*, vol. 92, pp. 80–92, Apr. 2023.
[20] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
[21] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.
[22] Z. Zhao et al., "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5906–5916.
[23] Z. Zhao et al., "Equivariant multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 25912–25921.
[24] Z. Huang, J. Li, N. Mao, G. Yuan, and J. Li, "DBEF-Net: Diffusion-based boundary-enhanced fusion network for medical image segmentation," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124467.
[25] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-Fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, vol. 32, pp. 5705–5720, 2023.
[26] M. Zhang, L. Ou, C. Zhang, K. Luo, S. Liao, and C. Tang, "Sub-6G domain-adaptive non-contact sensing using transformer and prototypical neural network," *IEEE Trans. Cogn. Commun. Netw.*, early access, Mar. 17, 2025, doi: 10.1109/TCCN.2025.3551814.
[27] J. Li et al., "MetaFormer-based lightweight neural network for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 275–279, Feb. 2025.
[28] M. Zhang et al., "IBSS: Intelligent behavior sensing system based on comparative language-CSI learning," *IEEE Sensors J.*, vol. 25, no. 13, pp. 25570–25584, Jul. 2025.
[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016, *arXiv:1409.0473*.
[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[32] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," 2015, *arXiv:1410.8516*.
[33] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," 2017, *arXiv:1605.08803*.
[34] L. Ardizzone et al., "Analyzing inverse problems with invertible neural networks," 2019, *arXiv:1808.04730*.
[35] S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone, and U. Köthe, "BayesFlow: Learning complex stochastic models with invertible neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1452–1466, Apr. 2022.
[36] D. He, W. Li, G. Wang, Y. Huang, and S. Liu, "MMIF-INet: Multimodal medical image fusion by invertible network," *Inf. Fusion*, vol. 114, Feb. 2025, Art. no. 102666.
[37] C. Pan et al., "DANet: A dual-branch framework with diffusion-integrated autoencoder for infrared–visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–13, Mar. 2025.
[38] B. Verbeken and M.-A. Guerry, "Attainability for Markov and semi-Markov chains," *Mathematics*, vol. 12, no. 8, p. 1227, 2024.
[39] A. Verma, T. Badal, and A. Bansal, "Advancing image generation with denoising diffusion probabilistic model and ConvNeXt-V2: A novel approach for enhanced diversity and quality," *Comput. Vis. Image Underst.*, vol. 247, Oct. 2024, Art. no. 104077.
[40] H. Cao, Y. Tian, Y. Liu, and R. Wang, "Water body extraction from high spatial resolution remote sensing images based on enhanced U-net and multi-scale information fusion," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 16132.
[41] X. Chen, S. Xu, S. Hu, and X. Ma, "MGFA: A multi-scale global feature autoencoder to fuse infrared and visible images," *Signal Process., Image Commun.*, vol. 128, Oct. 2024, Art. no. 117168.
[42] K. Luo, L. Ou, M. Zhang, S. Liao, and C. Zhang, "A dictionary learning based unsupervised neural network for single image compressed sensing," *Image Vis. Comput.*, vol. 151, Nov. 2024, Art. no. 105281.
[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.
[44] A. Toet and M. A. Hogervorst, "Progress in color night vision," *Opt. Eng.*, vol. 51, no. 1, 2012, Art. no. 10901.
[45] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vols. 83–84, pp. 79–92, Jul. 2022.
[46] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12484–12491.
[47] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3496–3504.
[48] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, 2021.
[49] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma, "Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior," *Inf. Fusion*, vol. 110, Oct. 2024, Art. no. 102450.
[50] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, 2008, Art. no. 23522.
[51] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," *Inf. Fusion*, vol. 2, no. 3, pp. 169–176, 2001.
[52] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Math. Program.*, vol. 162, pp. 83–112, Mar. 2017.
[53] J.-M. Sung, D.-C. Kim, B.-Y. Choi, and Y.-H. Ha, "Image thresholding using standard deviation," in *Proc. 7th Image Process., Mach. Vis. Appl.*, 2014, pp. 182–188.
[54] A. G. Asuero, A. Sayago, and A. González, "The correlation coefficient: An overview," *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41–59, 2006.
[55] A. Pareto, "A new look at the correlation coefficient: Correlation as the difference-sum ratio of SSEs," *Commun. Statist.-Theory Methods*, vol. 52, no. 9, pp. 2852–2859, 2023.
[56] A. Amankwah and C. Aldrich, "Multiresolution image registration using spatial mutual information," in *Proc. Oceans*, 2012, pp. 1–4.
[57] T.-Y. Kuo, C.-M. Tsai, C.-P. Chuang, and S.-J. Chuang, "Image quality assessment with visual sensitivity," in *Proc. Int. Conf. Inform., Electron. Vis. (ICIEV)*, 2015, pp. 1–4.
[58] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. Int. Conf. Image Process.*, 2003, pp. 3–173.
[59] Z.-S. Xiao, "An image fusion assessment metric based on multi-scale structure similarity," *Appl. Mech. Mater.*, vol. 215, pp. 674–678, Nov. 2012.
[60] G. Jocher et al. "ultralytics/yolov5: V6. 0-YOLOv5n'Nano'models, Roboflow integration, TensorFlow export, OpenCV DNN support." Zenodo. 2021. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2021zndo...5563715J/abstract

Chengyi Pan received the B.E. degree in computer science and technology from Xiamen University of Technology, Xiamen, China, in 2023. He is currently pursuing the master's degree with the School of Software, Yunnan University, Kunming, China.

His research interests include image fusion, computer vision, and deep neural networks.

Qian Jiang (Member, IEEE) received the B.S. degree in thermal energy and power engineering and the M.S. degree in power engineering and engineering thermo-physics from Central South University, Changsha, China, in 2012 and 2015, respectively, and the Ph.D. degree in communication and information systems from Yunnan University, Kunming, China, in 2019.

She was a Postdoctoral Fellow with the School of Software, Yunnan University from 2019 to 2021, where she is currently an Associate Professor with the School of Software. Her research interests include deep neural networks, fuzzy set theory, bio-informatics, image processing, and information fusion.

Huangqimei Zheng received the B.E. degree in computer Science and technology from Xiamen University of Technology, Xiamen, China, in 2023. She is currently pursuing the master's degree with the School of Software, Yunnan University, Kunming, China.

Her research interests include multispectral sensing, image fusion, and deep neural networks.

Hongyue Huang received the B.E. degree in communication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2015, the M.S. degree in computer science from Technische Universität Berlin, Berlin, Germany, in 2018, and the Ph.D. degree in computer science from Vrije Universiteit Brussel, Ixelles, Belgium, in 2021.

From 2022 to 2024, he served as a Postdoctoral Researcher with the School of Computer Science, Peking University, Beijing, China. He is currently a Lecturer with the School of Software (School of Artificial Intelligence), Yunnan University, Kunming, China. His research interests centered on visual information processing.

Xin Jin (Senior Member, IEEE) received the B.S. degree in electronics and information engineering from Henan Normal University, Xinxiang, China, in 2013, and the Ph.D. degree in communication and information systems from Yunnan University, Kunming, China, in 2018.

He was a Postdoctoral Fellow with the School of Software, Yunnan University from 2018 to 2020, where he is an Associate Professor with the School of Software. His research interests include pulse coupled neural networks and its applications, image processing, information fusion, optimization algorithm, and fuzzy set theory.

Keqin Li (Fellow, IEEE) received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 1985, and the Ph.D. degree in computer science from the University of Houston, Houston, TX, USA, in 1990.

He is a SUNY Distinguished Professor with the State University of New York, New Paltz, NY, USA, and a National Distinguished Professor with Hunan University, Changsha, China. He has authored or co-authored more than 1130 journal articles, book chapters, and refereed conference papers. He holds nearly 80 patents announced or authorized by the Chinese National Intellectual Property Administration.

Dr. Li has been among the World's Top Few Most Influential Scientists in Parallel and Distributed Computing Regarding Single-Year Impact (Ranked #2) and Career-Long Impact (Ranked #4) based on a composite indicator of the Scopus citation database since 2020. He is listed in Scilit Top Cited Scholars from 2023 to 2024 and is among the top 0.02% out of over 20 million scholars worldwide based on top-cited publications. He is listed in ScholarGPS Highly Ranked Scholars from 2022 to 2024 and is among the top 0.002% out of over 30 million scholars worldwide based on a composite score of three ranking metrics for research productivity, impact, and quality in the recent five years. He received the IEEE TCCLD Research Impact Award from the IEEE CS Technical Committee on Cloud Computing in 2022 and the IEEE TCSVC Research Innovation Award from the IEEE CS Technical Community on Services Computing in 2023. He won the IEEE Region 1 Technological Innovation Award (Academic) in 2023. He was a recipient of the 2022–2023 International Science and Technology Cooperation Award and the 2023 Xiaoxiang Friendship Award of Hunan Province, China. He is a member of the SUNY Distinguished Academy. He is an AAAS Fellow, an AAIA Fellow, an ACIS Fellow, and an AIIA Fellow. He is a member of the European Academy of Sciences and Arts and Academia Europaea (Academician of the Academy of Europe).

Wei Zhou (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently a Full Professor with the Software School, Yunnan University, Kunming, China. He has hosted several National Natural Science Foundation projects. His current research interests include distributed data-intensive computing and bioinformatics.

Prof. Zhou won the Wu Daguan Outstanding Teacher Award of Yunnan University in 2016. He was selected into the Youth Talent Program of Yunnan University in 2017. He is currently a Fellow of China Communications Society and a member of Yunnan Communications Institute and the Bioinformatics Group, Chinese Computer Society.