IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2023, accepted 4 January 2024, date of publication 9 January 2024,
date of current version 17 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3351774

RESEARCH ARTICLE

High Efficient Spatial and Radiation Information Mutual Enhancing Fusion Method for Visible and Infrared Image

ZONGZHEN LIU[1,2,3,4,5], YUXING WEI[2,3,4,5], GELI HUANG[1], CHAO LI[1], JIANLIN ZHANG[2,3,4,5], MEIHUI LI[2,3,4,5], DONGXU LIU[2,3,4,5], AND XIAOMING PENG[1]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]National Key Laboratory of Optical Field Manipulation Science and Technology, Chinese Academy of Sciences, Chengdu 610209, China
[3]Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu 610209, China
[4]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
[5]University of Chinese Academy of Sciences, Beijing 101408, China

Corresponding authors: Yuxing Wei (yuxing_wei_ioe5@163.com) and Xiaoming Peng (pengxm@uestc.edu.cn)

This work was supported in part by the Frontier Research Fund of Institute of Optics and Electronics, China Academy of Sciences, under Grant C21K005; and in part by the National Natural Science Foundation of China under Grant 62101529.

ABSTRACT Visible and infrared image fusion is an important image enhancement technique that aims to generate high-quality fused images with prominent targets and rich textures in extreme environments. However, most of the current image fusion methods have poor visual perception of the generated fused images due to severe degradation of texture details of the visible light images in scenes with extreme lighting, which seriously affects the application of subsequent advanced vision tasks such as target detection and tracking. To address these challenges, this paper bridges the gap between image fusion and advanced vision tasks by proposing an efficient fusion method for mutual enhancement of spatial and radiometric information. First, we design the gradient residual dense block (LGCnet) to improve the description of fine spatial details in the fusion network. Then, we developed a cross-modal perceptual fusion (CMPF) module to facilitate modal interactions in the network, which effectively enhances the fusion of complementary information between different modalities and reduces redundant learning. Finally, we designed an adaptive light-aware network (ALPnet) to guide the training of the fusion network to facilitate the fusion network to adaptively select more effective information for fusion under different lighting conditions. Extensive experiments show that the proposed fusion approach has competitive advantages over six current state-of-the-art deep-learning methods in highlighting target features and describing the global scene.

INDEX TERMS Image fusion, cross-modal, light perception, visible and infrared image, mutual enhancement.

I. INTRODUCTION

Due to the advancement of imaging devices and analytical techniques, multi-modal visual data have become pervasive. However, in complex scenarios like low visibility, occlusion, and day-night transitions, a single-mode camera cannot generate images suitable for specific tasks, such as continuous scene monitoring. This is mainly because single-modality information cannot characterize the imaged scene efficiently and comprehensively [1]. Therefore, image fusion, as a viable alternative, has been extensively researched. It enables improved understanding of the correlation and interaction between information in images from various modalities. Visible and infrared image fusion, one of the most representative multi-modal image fusion techniques, merges the rich texture and high spatial resolution from visible images with the structural and thermal information from infrared images. Visible-infrared image fusion has found extensive

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

applications in military and security domains, including target detection [2] and tracking [3].

Over the past decade, numerous methods for fusing visible and infrared images have been introduced. These methods can be broadly classified into traditional approaches [4], [5], [6] and more recent deep learning-based ones. The traditional approaches can be more specifically classified into five categories: 1) methods based on multi-scale transforms (MST) [4], [7], [8], [9], [10], [11], 2) methods based on sparse representations (SR) [12], [13], [14], 3) methods based on subspaces [15], [16], 4) methods based on low-rank representations (LRR) [17], [18], [19], [20], [21], and 5) miscellaneous methods that do not fit into the four categories above [22], [23]. While still applicable to specific tasks, the traditional methods have become outdated for two reasons. Firstly, conventional methods are often intricate and, at the same time, pay insufficient attention to the distinguishing characteristics of visible and infrared images. Secondly, crafting rules manually for the adaptive fusion of visible and infrared information across various scenes [24].

Mainstream deep learning-based methods can be broadly classified into four groups: auto-encoder (AE)-based [25], [26], [27], [28], [29], [30], convolutional neural network (CNN)-based [31], [32], [33], transformer-based [34], [35], [36] and generative adversarial network (GAN)-based [37], [38], [39], [40]. While these approaches demonstrate strong performance, their fusion speed is relatively slow. Furthermore, the majority of deep learning-based algorithms initially concatenate the source images before inputting them into a single-path network, which lacks individual feature extraction branches for each input image. There are also fusion algorithms that operate at the target level [41], [42]. In these methods, semantic segmentation results from fused images are used to enhance the fusion of information from various targets, enabling the fused images to effectively preserve the semantic content from both infrared and visible light sources. However, when the visible image is affected by severe light pollution, it can introduce errors in the model's segmentation results, leading to poor image fusion.

To address these issues, we propose a highly efficient spatial and radiation information mutual enhancing fusion method. Specifically, to meet the real-time requirement of advanced vision tasks, we design a lightweight network based on gradient residual dense blocks (LGCnet). LGCnet enables feature reuse through the main dense stream and improves the characterization of fine-grained details through the residual gradient stream. Subsequently, we designed a light-aware network to guide the training of our model, with the aim of enabling our model, under different lighting conditions, to adaptively select valid information from infrared and visible images for fusion. Finally, we design a cross-modal perceptual fusion module (CMPF module) for fully extracting and fusing complementary information from visible and infrared images and enhancing the interaction between the two modalities while reducing redundant learning. Extensive experiments on three mainstream visible-infrared image fusion datasets—RoadScene [43], M3FD [14], and MSRS [44] —show that the proposed approach outperforms state-of-the-art competitors.

Our main contributions include the following three parts:

1) We propose an efficient mutual enhancement fusion method of spatial and radiometric information, which can provide high-quality fused images for subsequent vision tasks such as target detection and tracking while better meeting their real-time requirements.

2) We designed a light-aware network (ALPnet), which can adaptively select effective information from infrared and visible images for fusion under different lighting conditions, effectively minimizing the effect of light pollution on the fusion model.

3) We developed a cross-modal information fusion module (CMPF module), which effectively enhances the interaction between thermal radiation information in infrared images and rich texture information in visible images, and reduces redundant learning between modalities.

The rest of the paper is organized as follows. In Section II, we briefly describe and analyze the related work, including both traditional and deep-learning-based image fusion approaches. In Section III, we describe our proposed method in detail, including problem analysis, network architecture, and loss function. In Section IV, we first present the experimental configuration, data evaluation metrics, and detailed experimental details. Then, the experimental results of our proposed method on three mainstream visible-infrared image fusion datasets are presented and its performance is compared with several existing methods. Finally, we will summarize the conclusions and future work of this paper in Section V.

II. RELATED WORK

In this section, First, in Section A, we review two types of traditional image fusion methods. Then, in Section B, we describe and discuss deep-learning-based image fusion methods. Finally, in Section C, we introduce some of the work done by predecessors in enhancing image quality with lighting-aware technologies.

A. TRADITIONAL FUSION METHODS

Research interest in traditional methods mainly focuses on multi-scale transform and sparse representation.

Multi-scale transform methods decomposes each source infrared or visible image, into a base layer and multiple detail layers at various scales. The base layer is employed to manage the overall contrast of the fused image, whereas the detail layers are utilized to handle the detailed information within the fused image [4].

The sparse representation coefficients of the source images are derived from the acquired overcomplete dictionary. These coefficients are then fused according to an specified fusion rules. The acquired overcomplete dictionary is employed to reconstruct the fused image from the combined

FIGURE 1. The first and third rows represent the infrared images, the visible images, and the fused images, respectively. The second and fourth rows represent the feature maps that are corresponding to the infrared images, the visible images, and the fused images, respectively.

sparse representation coefficients. Notably, it should be noted that more intricate transformations and representations cannot fulfill the requirements for real-time image fusion [14]. Moreover, manually devised fusion rules fall short of integrating semantic information, which ultimately constrains the contribution of fusion to advanced visual tasks.

B. DEEP LEARNING-BASED FUSION METHODS

Autoencoder methods. Autoencoder methods, which involve neural networks comprising an encoder and a decoder, have gained prominence in image fusion in recent years. These methods typically begin with pre-training an auto-encoder to extract features and recover images, followed by feature fusion using conventional rules. For instance, Li and Wu [27] introduced DenseFuse, a novel image fusion network that incorporates dense blocks into both the encoder and decoder. Li et al. [26] adopted an encoder to extract multi-scale features from source images and a decoder based on nested connectivity architecture to reconstruct fused features, including spatial and channel attention. In another approach, Liu et al. [45] utilized two encoders to extract diverse intrinsic features from source images of varying modalities, culminating in the use of a unified decoder to obtain the fused image. This passage highlights the prevalence of auto-encoder networks in image fusion and outlines their application in various studies, detailing feature extraction, encoder and decoder structures, and fusion processes. Wang et al. [30] constructed a novel image fusion network based on an autoencoder, which combines a CNN and a transformer to capture both local and global features of the source image, effectively enhancing the contrast and gradient information of the fused image.

GAN methods. An increasing number of GAN-based fusion approaches have been proposed, leveraging GANs' suitability for image fusion due to their capacity for unsupervised distribution estimation. Ma et al. [38] initially introduced an adversarial framework to enhance texture structure preservation by pitting the fusion result against the visible image. However, this single adversarial mechanism can result in imbalanced fusion. To address this issue, they later introduced a dual discriminator conditional generative adversarial network (DDcGAN) [39] for image fusion, involving both visible and infrared images in the adversarial process. Notably, GANs with dual discriminators can be challenging to train. Subsequently, MFEIF [46] and AttentionFGAN [47] improved fusion performance by introducing feature attention blocks and multi-scale attention blocks in the model, respectively. However, these attention blocks struggled to extract and retain unique information from different source images.

Transformer methods. The transformer is widely used in fusion networks. Rao et al. [35] combine GAN and Transformer, using the transformer module and convolutional module as a generator in a GAN to generate images and different Transforms in the generator to ensure the image's spatial correlation and dimensional correlation. Then, the discriminator is used to ensure that the fused image can retain details in the visible image and essential information in the infrared image. Ma et al. [36] combine CNN with Swin-Transformer, using CNN to extract the local information from the image and Swin-Transformer to extract global information of the image. These information features are fused within domain and cross-domain. Mustafa et al. [34] extract shallow features by convolutional operations, and then introduce the transformer block in the feature extraction process to effectively capture the local and global relationships between complementary features, spaces, and channels. The graph-attention fusion block (GAFB) is used to improve the selectivity and effectiveness of the feature learning.

CNN methods. Leveraging the exceptional feature extraction capabilities of neural networks, deep learning has led to remarkable advances in image fusion. Early efforts involved the use of Convolutional Neural Networks (CNNs) for multi-focal image fusion. However, these networks were tailored exclusively for multi-focal image fusion. Subsequently, neural networks were employed to create weight maps or extract features for visible and infrared image fusion [43]. Building on these achievements, Zhang et al. developed a generalized image fusion framework based on a versatile network structure encompassing feature extraction, fusion, and image reconstruction layers [31], [32]. This approach is not limited to multi-focus image fusion; it is also suitable for infrared, visible light, medical image fusion, and other applications. Nonetheless, these neural networks lack

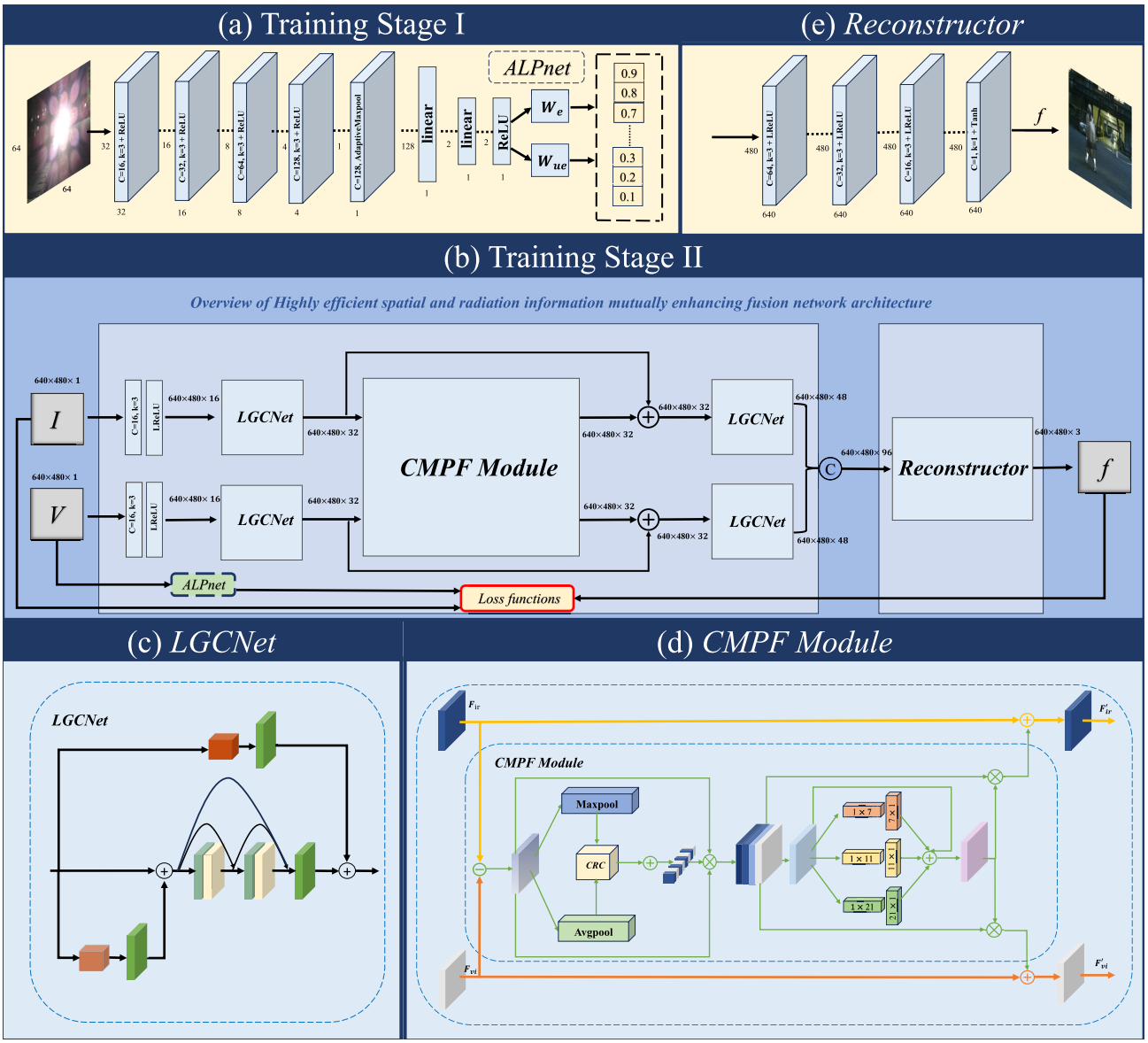


FIGURE 2. Overview of highly efficient spatial and radiation information mutually enhancing fusion network architecture.

specific training for image fusion, constraining their fusion performance.

C. LIGHTING-AWARE IMAGE ENHANCEMENT

Several research studies have studied the effect of lighting factors in the modeling process. In Figure 1, it can be seen that the image quality is impacted by different lighting conditions. In the case of sufficient lighting, visible images with adequate texture details are usually obtained, while in the case of insufficient illumination, the quality of the images captured by the sensor is severely degraded, and these images usually show reduced clarity, poorly defined edges, and poor overall visibility. To address these problems, some research has focused on image enhancement techniques, where researchers have attempted to utilize neural networks to estimate the reflectance and illumination distribution of an image in order to improve image enhancement under low-light conditions. RetinexNet [48] incorporates the Retinex theory [49] into its network architecture and designs a deep-learning network consisting of a decomposition module and an illumination tuning module. These techniques highlight important features in an image to improve clarity, making it more suitable for human or computer analysis and processing tasks [21]. Guan et al. proposed a multi-spectral pedestrian detection framework based on illumination awareness that combines illumination awareness and semantic segmentation [50]. MBNet employs a flexible and balanced optimization approach to improve the performance of the detector [51], which uses a light-aware feature alignment module to adaptively select complementary information from

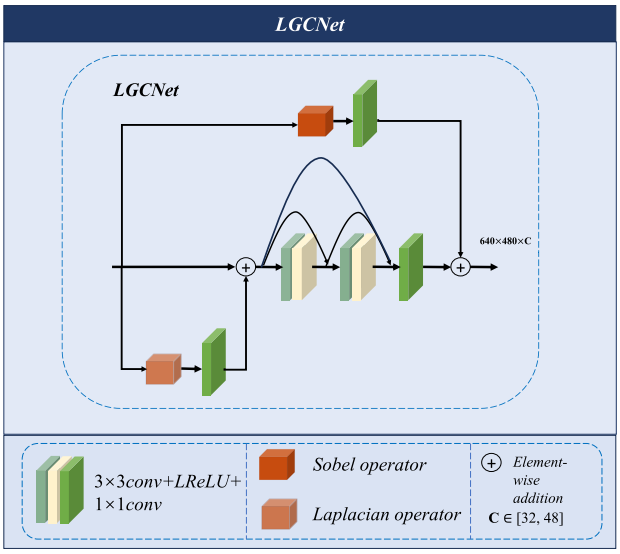


FIGURE 3. Deep feature extraction network (LGCnet).

both the visible and infrared images based on lighting conditions.

In our study, we designed a light-aware network to allow our fusion network to adaptively adjust the ratio of fusion information between visible and infrared images, thus effectively reducing the effect of illumination on the fused images. We can find from Figure 1 that in the case of low light intensity, the infrared image has more significant target information, while the detailed information of the visible image is seriously affected by the light. In the case of sufficient light, the visible image has more detailed information compared to the infrared image. For better image fusion, it is necessary to design a light-aware network to guide the generation of fused images.

III. PROPOSED FUSION METHOD

In this section, we describe the proposed visible-infrared image fusion network. First, in Section A, we overview the entire fusion network we have designed. Then, in Section B, we provided a detailed introduction to the working principle of the light-aware network. Then, in Sections C and D, we have provided a detailed introduction to how the LGCNet and the cross-modal perception fusion module operate. Finally, in Section E we design the loss functions used to train the proposed network.

A. NETWORK ARCHITECTURE

To implement a real-time speed for image fusion, we propose a visible and infrared image fusion network based on LGCnet we designed, as shown in Figure 2. Our fusion network consists of a feature extractor and an image reconstructor. The feature extractor contains four LGCnet blocks and one cross-modal fusion perception module (CMPF module). As a whole, the feature extractor is responsible for extracting features from the input images and, at the same time, enhancing the interaction between the two different modalities. Two parallel feature extraction streams, each for one modality, are implemented. At the same time, in our entire model architecture, we extensively use the Conv+LeakyReLU structure instead of the traditional Conv+BN+Activation structure. We experimentally found out that incorporating BN layers did not improve the performance of the proposed method further. Meanwhile, the extra BN layers would increase the computational load. An input image first undergoes a 3×3 convolution layer and an LReLU activation function for shallow feature extraction. Then, the results are passed through two LGCnet modules and a common cross-modal fusion perception module (CMPF module). The LGCnet module, shown in Figure 3, is designed for deeper feature extraction purposes. The cross-modal fusion perception module (CMPF module) shown in Figure 4, which enables our network to integrate more effective information in the feature extraction stage, mainly consists of the channel attention module *CA* and the spatial attention module *SA*. Then, we concatenate the visible and infrared features and feed them into the image reconstructor for feature aggregation and image reconstruction. The reconstructor, shown in Figure 2., The image reconstructor consists of three 3×3 convolutional layers and one 1×1 convolutional layer in series. All 3×3 convolutional layers use LReLU (Leaky Rectified Linear Unit) as the activation function, while the 1×1 convolutional layer uses the Tanh function as the activation function.

B. LIGHT-AWARE NETWORK

Given an infrared image I_{ir} and a visible image I_{vi}, a fused image I_f can be generated by feature extraction, fusion, and reconstruction. To this end, a light-aware loss is designed that reflects the above three steps.

Considering that light imbalance affects information distribution, we develop a light-aware network, as shown in Figure 2, to estimate the illumination of a visible image adaptively. Given a visible image I_{vi}, the process of light perception can be defined as:

$$\{W_e, W_{ue}\} = N_{ALP}(I_{vi}) \tag{1}$$

where N_{ALP} refers to the light-aware network, and W_e and W_{ue} represent the probabilities of sufficient and insufficient lighting, respectively. Both W_e and W_{ue} are non-negative. Since visible images have more useful information under sufficient lighting conditions and infrared images have more useful information under insufficient light conditions, the light conditions also directly affect the richness of the information contained in the images. Therefore, we guide the fusion of visible and infrared images by calculating the illumination probability of the current scene. When W_e is high, the visible image contributes more to the fused image. By contrast, when W_{ue} is high, the infrared image contributes more to the fused image. Hence, W_e and W_{ue} control the degree of contribution of each modality to the fused image.

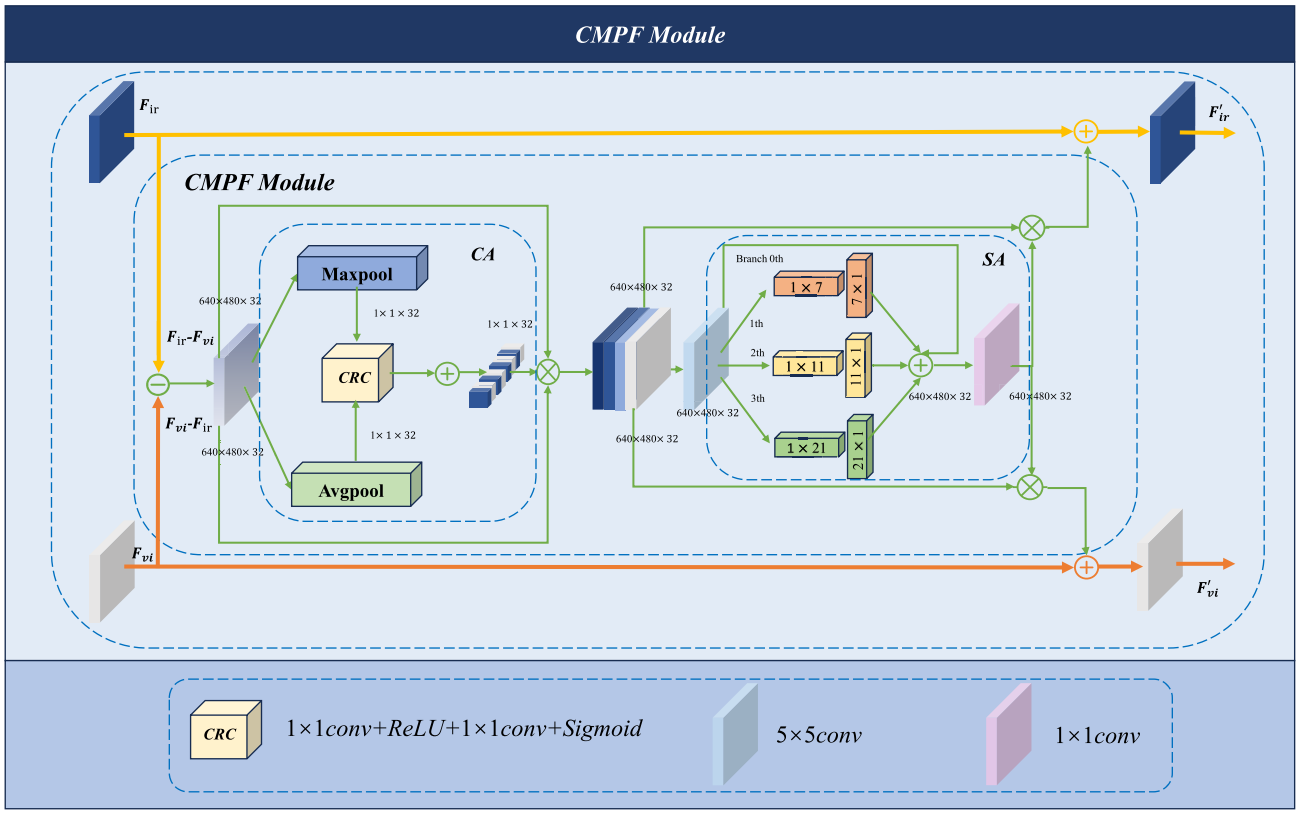


FIGURE 4. Cross-modal perceptual fusion module (CMPF module).

C. THE LGCNET

The LGCnet module consists of two 3×3 convolution layers, two LReLU activation functions, three 1×1 convolution layers, a Sobel gradient operator, and a Laplacian operator. Both operators are applied to a feature map in a depthwise convolution manner, namely, each channel of the feature map is convolved with one of the operators individually. The sobel filtering involves two 3×3 masks, with their structures given as follows: $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ and $\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$. They are used to extract horizontal and vertical edges from each channel of the feature map, respectively. The Laplacian operator has one 3×3 mask, whose structure is $\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$. The input to both the Sobel and Laplacian operators is appropriately padded to ensure the output of the filtering operation has the same size as that of the input. Though both the Sobel and Laplacian operators are designed using image derivatives, they are different in that the former is first-order-derivative-based while the latter is second-order-derivative-based. In practice, the Sobel operator is mainly used to extract edges from a grayscale image while the Laplacian operator is usually used to highlight high-frequency details in the image. Due to their differences, we can expect to have a complementary power from each operator by combing them into the proposed LGCNet.

However, as a feature map contains multiple channels, it would be desirable to introduce interactions between the results of Laplacian or Sobel filtering of each channel. For this purpose, a fully-connected network can be used. The fully-connected network can be viewed as a linear transform that is optimally learnt to make the individual channels of the Laplacian or Sobel filtering results to interact with each other. Here, we use a 1×1 convolution layer to act as the intended fully-connected network. Further, the 1×1 convolution layer keeps the number of channels intact for the input and output of the Laplacian and Sobel filtering operations. On the other hand, we were also inspired by ResNet to follow the Laplacian and Sobel filtering operations with a 1×1 convolution layer. Please note that in ResNet, the original input is passed through a convolution+ReLU block before being added back to itself. The feature order might be disrupted, but this disruption might be desirable because it results in necessary interactions between the channels of the feature map.

D. FEATURE-BASED CROSS MODAL ENHANCEMENT

We design a lightweight real-time fusion network. Specifically, we develop a feature extraction network E_f, formulated

as:

$$\{F_{vi}, F_{ir}\} = \{E_f(I_{vi}), E_f(I_{ir})\} \tag{2}$$

where F_{vi} and F_{ir} represent the features of the visible image and infrared image extracted by E_f, respectively. We formulate the common and complementary parts of F_{vi} and F_{ir} as follows:

$$F_{vi} = \frac{F_{vi} + F_{vi}}{2} + \frac{F_{ir} - F_{ir}}{2} = \frac{F_{vi} + F_{ir}}{2} + \frac{F_{vi} - F_{ir}}{2}, \tag{3}$$

$$F_{ir} = \frac{F_{ir} + F_{ir}}{2} + \frac{F_{vi} - F_{vi}}{2} = \frac{F_{ir} + F_{vi}}{2} + \frac{F_{ir} - F_{vi}}{2}, \tag{4}$$

where the common part represents the common features of visible and infrared images, while the complementary part reflects their complementary features. In addition, we propose a cross-modal fusion perception module (CMPF module) to enhance the common and complementary information of visible infrared images, defined as:

$$F_c(F) = CA(F) \otimes F \tag{5}$$

$$\widehat{F}(F) = SA(Conv_{5\times5}(F_c(F))) \otimes F_c(F) \tag{6}$$

$$CA(F) = \delta(CRC(Maxpool(F)) + CRC(Avgpool(F))) \tag{7}$$

$$CRC(F) = \delta(Conv_{1\times1}(ReLU(Conv_{1\times1}(F)))) \tag{8}$$

$$SA(F) = Conv_{1\times1}(\sum_{i=0}^{3} Branch_i(DwConv(F))) \tag{9}$$

$$F'_{ir} = \widehat{F}(F_{ir} - F_{vi}) + F_{ir} \tag{10}$$

$$F'_{vi} = \widehat{F}(F_{vi} - F_{ir}) + F_{vi} \tag{11}$$

where $\otimes$ represents element-wise multiplication, $CA(\cdot)$ represents the channel attention module, $SA(\cdot)$ represents the spatial attention module, and $\delta(\cdot)$ is the Sigmoid function, DwConv represents depth-wise convolution, and $Branch_i$, $i \in \{0, 1, 2, 3\}$ represents the *ith* branch. $Branch_0$ is the identity connection. The CMPF module uses the difference of F_{ir} and F_{vi}, $F_{vi} - F_{ir}$ and $F_{ir} - F_{vi}$, as input. This module enables $F_{vi} - F_{ir}$ and $F_{ir} - F_{vi}$ to learn complementary information from each other. It mainly consists of the channel attention module CA and the spatial attention module SA. The workflow of the CMPF module can be summarized by (Eqs. 5, 6, 7, 8, 9, 10, 11) and described in Figure 4. Since $F_{vi} - F_{ir}$ and $F_{ir} - F_{vi}$ are processed in a same way, for simplicity we will use F to represent either of them. F is parallelly pooled using two different pooling strategies, adaptive average pooling and adaptive max pooling. The results are then sent through the CRC blocks (Eq. 8) before being aggregated by the CA module (Eq. 7). Since the two pooling strategies are complementary with each other, as proved in CBAM [52], they are both employed in the CMPF module. The output of the CA module is used as channel-wise weights for F to obtain $F_c(F)$ (Eq. 5). Next, $F_c(F)$ passes through a convolution layer before entering into the SA module, the output of which is in turn used as channel-wise weights for $F_c(F)$ (Eq. 6). The SA module contains a multi-scale structure $Branch_i$ (i=0,1,2,3), which generates multi-scale attention maps with refined complementary information (Eq. 9). Depthwise convolution in the SA module is responsible for capturing inter-channel spatial relationships between features. At the same time, it also reduces computational complexity. Finally, the output of the SA module is a feature map which contains complementary information from both the infrared and visible images; it is thus added to the respective infrared and visible feature maps, F_{ir} and F_{vi}, to obtain enriched results F'_{ir} and F'_{vi} (Eqs. 10 and 11). After the feature extraction and encoding of the visible and infrared images, we decode the features of the visible and infrared images to get the final fused image, formulated as:

$$F_f = \mathrm{C}(F_{vi}, F_{ir}) \tag{12}$$

$$I_f = D(F_f) \tag{13}$$

where $\mathrm{C}(\cdot)$ means the concatenation of the visible and infrared features in the channel direction while $D(\cdot)$ stands for feature reconstruction.

E. LOSS FUNCTIONS

ALPNet is a simple binary classification network consisting of four 3 × 3 convolutional layers, four ReLU (Rectified Linear Unit) activation functions, a maximum pooling layer, two linear layers, and one Relu activation function, as shown in Figure 2. The ALPNet network was trained separately (Training stage I, Figure 2(a)) before the whole image fusion network was trained (Training stage II, Figure 2(b)). The output of the ALPNet network is (W_e, W_{ue}) (Eq. 1), where W_e and W_{ue} represent the probabilities of sufficient and insufficient lighting, respectively. (W_e, W_{ue}) is used to weigh the illumination loss function L_{illum} (Eq. 15) during the training of the whole image fusion network. Please note that the ALPNet network is not used in the inference stage. We use the cross-entropy loss formulated below to guide the training of the light-aware network:

$$L_{ALP} = -z \log \sigma(y) - (1 - z) \log(1 - \sigma(y)) \tag{14}$$

where z refers to the label of the input image, and $y = \{W_e, W_{ue}\}$ refers to the output of the light-aware network, and $\sigma(\cdot)$ represents the softmax function

To enable our fusion network to better adaptively fuse the effective information from a pair of visible and infrared images according to the lighting conditions of the scene, we propose the following illumination loss function:

$$L_{\text{illum}} = W_{ue} \cdot L_{\text{int}}^{ir} + W_e \cdot L_{\text{int}}^{vi} \tag{15}$$

where L_{int}^{vi} and L_{int}^{ir} represent the intensity loss functions for the visible and infrared images, respectively.

Further, we define the following intensity loss function below, which measures the pixel-level difference between the

fused image and the source images.

$$L_{\text{int}}^{ir} = \frac{1}{HW}||I_f - I_{ir}||_1 \tag{16}$$

$$L_{\text{int}}^{vi} = \frac{1}{HW}||I_f - I_{vi}||_1 \tag{17}$$

where H and W represent the height and width of the input image, respectively, and $||\cdot||_1$ represent the $L1$ distance. The intensity distribution of the fused image should be consistent with the different source images under different illumination conditions. Therefore, we use the illumination-aware weights, W_{ue} and W_e, to adjust the intensity of the fused images. Further, an auxiliary intensity loss is introduced to retain an optimal intensity distribution for the fused image, which can be expressed as

$$L_{\text{ints}} = \frac{1}{HW}||I_f - max(I_{ir}, I_{vi})||_1, \tag{18}$$

where $max(\cdot)$ indicates that the maximum value is selected. To explain why we use multiple different supervision values for the reconstruction of the fused image, consider a night-time scenario, in which the visible image is dark while the infrared image is bright because it captures the thermal emissions of the objects of interest in the scene. Concretely, let us suppose that for a same point on an object, the pixel value for this point on the visible image is 0.1, and the corresponding pixel value on the infrared image is 0.5. In this case, we would desire the corresponding pixel on the fused image to take the value of 0.5 instead of 0.1. The reason for this choice is because we would prefer brighter pixels to present in the fused image, as they are more likely to be associated with objects of interest. This way, the two pixel values of 0.1 and 0.5 are competing against each other to appear in the fused image. Thus, in a sense, the multiple different supervision values are competitive rather than redundant. A similar rule is adopted for a day-time scenario. By combining both scenarios, we have Eq. (18) for the reconstruction of the fused image.

With a same philosophy, to force the fused image to preserve sharp details of the source images, we introduce a texture loss function:

$$L_{\text{texture}} = \frac{1}{HW}|||\nabla I_f| - max(|\nabla I_{ir}|, |\nabla I_{vi}|)||_1, \tag{19}$$

where ∇ is the gradient operator intended to measure the texture information of the image. $|\cdot|$ represents the de-absolute operation.

Finally, our loss function for this network can be formulated as a weighted combination of light perception loss L_{illum}, auxiliary intensity loss L_{ints}, and texture L_{texture}:

$$L_{\text{fusion}} = \lambda_1 L_{\text{ints}} + \lambda_2 L_{\text{texture}} + \lambda_3 L_{\text{illum}}, \tag{20}$$

where λ_1, λ_2 and λ_3 are hyperparameters to balance the influence of the three losses. By incorporating the three losses into the whole loss function, the proposed network excels in generating a high-quality fused image that retains the optimal intensity distribution under the guidance of the light loss and auxiliary intensity loss. Under the guidance of texture loss, the fused image inherits the rich details of the original image.

IV. EXPERIMENTAL VALIDATION

We detail the experiments in this section. First, in Section A, we introduce the experimental configurations and details. Then, in Section B, we present the relevant evaluation metrics. Next, in Section C, we describe the implementation details of the training procedure of the proposed network. In Section D, we present the comparative experiments. To illustrate the generalization capability of the proposed approach, in Section E, we conducted generalization experiments, followed by Section F, in which we analyzed the performance of different fusion methods on the target detection task. Then, in Section G, we conducted efficiency comparisons with several other methods. We conclude this section with ablation experiments in Section H.

A. EXPERIMENTAL CONFIGURATIONS

In order to comprehensively evaluate the performance of our fusion algorithm, we selected 300, 50, and 30 pairs of images on the MSRS [44], RoadScene [43], and M3FD [14] datasets for qualitative and quantitative analysis of the proposed method. We selected six SOTA methods for comparison, namely DenseFuse [27], IFCNN [31], U2Fusion [43], TarDAL [14], SEDRFuse [53], and RFN-NEST [26].

B. EVALUATION METRICS

We chose eight evaluation metrics to quantitatively assess the performance of different fusion methods. These eight evaluation metrics are, entropy (EN) [54], spatial frequency (SF) [55], standard deviation (SD) [56], including average gradient (AG) [57], visual information fidelity (VIF) [58], structural similarity (SSIM) [59], fusion artifacts ($N^{ab/f}$) [60] and mutual information (MI) [61]. EN evaluates the amount of information contained in the fused image from an information theoretic perspective. SD measures the distribution and contrast of the fused image from a statistical point of view. AG reflects the richness of the texture information of the image. SF reflects the richness of the details of the fused image. SSIM measures the degree of similarity between images. $N^{ab/f}$ measures the robustness of a given method to artifacts or noise added in the fusion process. A lower value of $N^{ab/f}$ indicates a better performance. VIF measures the fidelity of the information from the point of view of human visual perception. MI measures the amount of information transferred from the source image to the fused image. The higher the results of SF, EN, SD, AG, MI, SSIM, and VIF, the better the performance of the fusion algorithm.

C. IMPLEMENTATION DETAILS

Our model training is divided into two stages; in the first stage, we train the light-aware network. First, we collect 29,853 images under adequate lighting conditions and 26,211 images under inadequate lighting conditions from the MSRS

dataset, and the size of the images is 64×64. The batch size is set to b, the train steps in one epoch are set as s, and it takes M epochs to train a model. Specifically, we set the batch size of the light-aware network to $b^1 = 128$, $M_1 = 100$, and $s_1 = 438$. In the second step, we choose the MSRS dataset to train our real-time fusion network; in the MSRS dataset, the training set contains 1,083 pairs of visible and infrared images, and the test set contains 361 pairs of images. Before feeding the training set into the fusion network, first, we transform the color space of the input images into the form of Ycbcr, then we fused the Y channel of the input images, and when the image fusion is complete, we transformed the fused image into the RGB color space by combining the cb and cr channels of the visible images. We set the parameters of the fusion network to $b^2 = 8$, $M_2 = 4$, $s_2 = 2700$, with an initial learning rate of 0.001, and update the relevant parameters with the Adam optimizer, as well as setting the hyper-parameters $\lambda_1 = 1$, $\lambda_2 = 10$, and $\lambda_3 = 0.01$ in the L_{fusion} formula. All our algorithms are implemented on the Pytorch platform, and all experiments are implemented on a server containing eight NVIDIA 3090 GPUs and an Intel(R) Xeon(R) Platinum 8375C CPU@2.90GHz.

D. COMPARATIVE EXPERIMENT

In order to fully evaluate the performance of our fusion algorithm, we compared our GACnet model with six other SOTA models on the MSRS dataset. Qualitative and quantitative analyses of the fusion results of different methods are given below.

1) QUALITATIVE RESULTS

The visualization results from four pairs of typical images on the MSRS dataset are shown in Figure 5. Our algorithms have a significant advantage, which is demonstrated by the fact that our fusion algorithms do a good job of fusing the limited visual detail information in the visible images with the salient target information in the infrared images in the poorly illuminated scenes. As in the first column of Figure 5, both TarDAL and SEDRFuse algorithms fail to clearly display the information of the person in the green box in the low-light scene. In the red box, the DenseFuse, RFN-NEST, SEDRFuse, TarDAL, and U2Fusion algorithms fail to fuse the salient target information in the infrared image well, and only IFCNN and our algorithm solve this problem well. In the fourth column of Figure 5, U2Fusion and DenseFuse have a hard time seeing the zebra line in the red box, while in RFN-NEST and SEDRFuse algorithms, the detail information of the zebra line is very blurred, and does not highlight the texture information of the zebra line well. In TarDAL and IFCNN algorithms, the detailed information in the infrared and visible images is not well fused, the thermal radiation in the crosswalk is more serious, and obvious black spots appear; only our algorithm solves all the above problems well.

Under sufficient lighting conditions, the visible image has rich texture information; we prefer the fused image to contain more information contributed from the visible image, while the thermal radiation information in the infrared image is used as a supplement to the visible information. In column 2 of Figure 5, the DenseFuse, SEDRFuse, U2Fusion, and TarDAL algorithms cannot clearly see the tree branches in the red box, while the RFN-NEST algorithm suffers from severe light pollution, and only IFCNN and our algorithms are able to see the texture information of the tree branches well. In column 3 of Figure 5, the DenseFuse, TarDAL, and U2Fusion algorithms cannot clearly display the detailed information about the headlights in the red box. SEDRFuse and IFCNN suffer from large light pollution. Only RFN-NEST and our algorithm completely retain the texture information of the visible images. From these images, our fusion algorithm is able to adaptively select effective information from visible and infrared images for fusion under different lighting conditions, which not only has comprehensive scene information but also preserves the rich contrast information and texture details of the target region.

2) QUANTITATIVE RESULTS

The quantitative results of the eight evaluation metrics for the 50 image pairs in the MSRS dataset are shown in Figure 6. Meanwhile, in order to quantitatively compare the results, we have ranked the different methods for the eight evaluation metrics as shown in Table 1. From the figure, we can find that our algorithm has a significant advantage in five evaluation metrics, namely, EN, SD, VIF, SSIM, and MI, and only lags behind the IFCNN algorithm in three metrics, namely, AG, $N^{ab/f}$ and SF. From these results, we show that our method can deliver more information from the source image to the fused image, retain useful global descriptive semantic information, ignore the redundant information in the target region, highlight the target region while retaining the rich scene information, and produce satisfactory fusion results that are also more in line with the human visual system. These excellent performances are attributed to our proposed illumination loss function and CMPF module.

E. GENERALIZATION EXPERIMENT

An important aspect of a given image fusion method is its generalization performance. Therefore, we conducted generalization experiments on the RoadScene and M3FD datasets to validate the generalization ability of our model. The following describes the quantitative and qualitative results of testing our model and six other SOTA models on these two datasets.

1) QUALITATIVE RESULTS

The qualitative results of the different algorithms on the M3FD dataset are shown in Figure 7. The IFCNN, DenseFuse, SEDRFuse, and U2Fusion algorithms lack the illumination information in the visible image and fail to retain the high contrast information of the visible image. The TarDAL algorithm fails to retain the rich texture detail

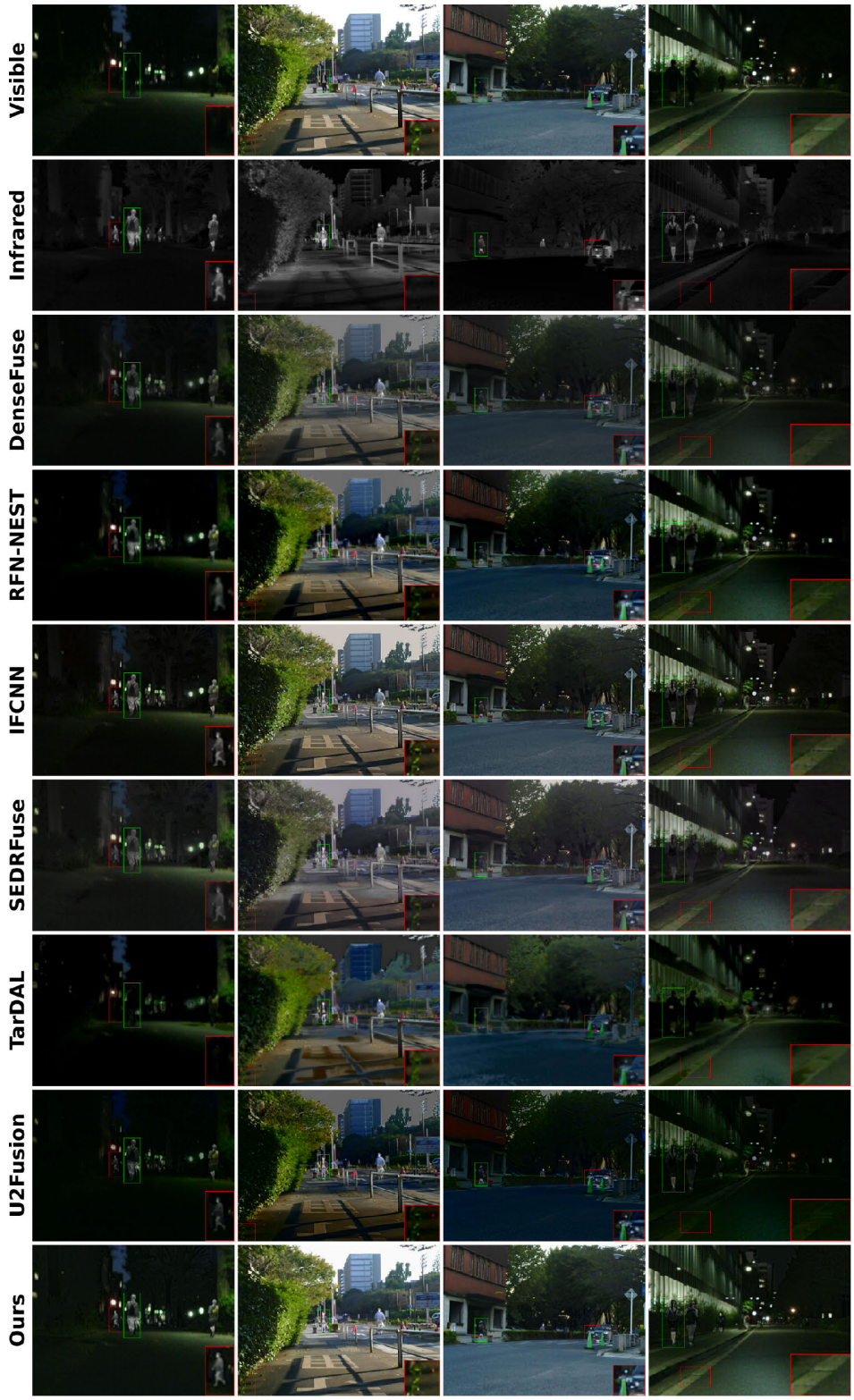


FIGURE 5. Qualitative comparison of GACnet in MSRS dataset with other SOTA six methods. For visualization purposes, we select a detail region from each image and place its zoom-in at the bottom of the image (in a red box). We also select a saliency region from each image and highlight it in a green box.

information of the visible image. Only our algorithm and the RFN-NEST algorithm succeed in preserving the saliency intensity of the target and the texture details of the visible image.

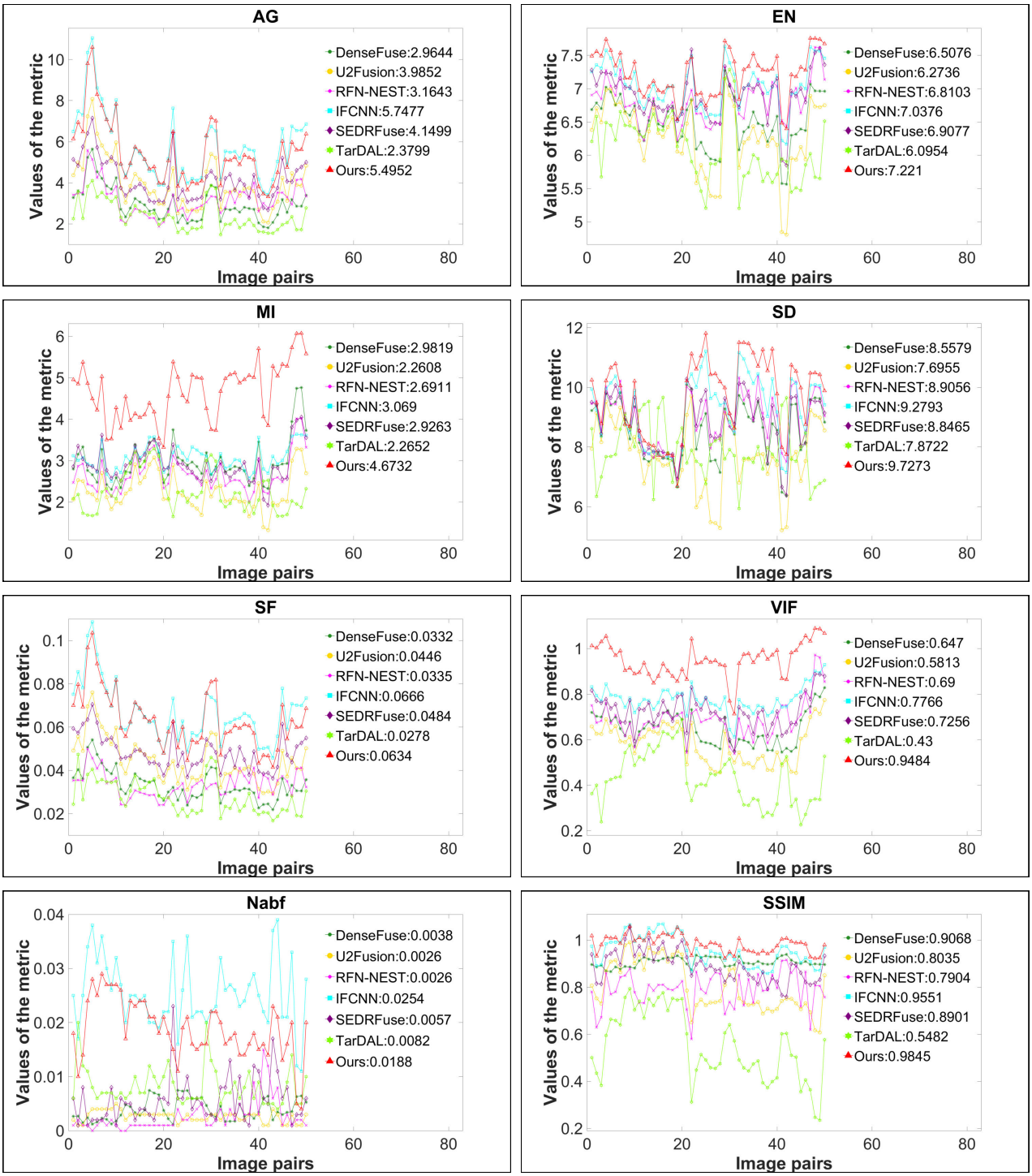


FIGURE 6. In order to clearly reveal the performance of the proposed approach and the six comparison methods, we only show the results of evaluating the eight metrics based on 50 pairs of images from the MSRS dataset. The higher the values of AG, EN, MI, SD, SF, VIF, and SSIM, the lower the values of $N^{a/bf}$, the better the performance of the model algorithm.

The qualitative results of the different algorithms on the RoadScene dataset are shown in Figure 8. The DenseFuse and IFCNN algorithms are heavily contaminated by visible light. The SEDRFuse algorithm retains too much information about the thermal radiation of the infrared image. The RFN-NEST, TarDAL, and U2Fusion fail to retain the information about the textural details of the visible image, and the image's textural details are blurred severely. Only our algorithm, which is able to adaptively select the useful information of visible and infrared images for fusion

TABLE 1. Experimental results of the MSRS dataset. The experiments were conducted three times. Each time 100 image pairs were randomly selected from the MSRS dataset to test the methods. The mean and standard deviation of the three experiments are given in the table, where the red markers represent the best results, the blue markers represent the second best results, and the green markers represent the third best results.

Model	DenseFuse	IFCNN	SEDRFuse	RFN-NEST	U2Fusion	TarDAL	Ours
AG	2.3874±0.1247	4.5945±0.2534	3.5244±0.2517	2.6204±0.2051	3.2805±0.2525	2.1329±0.5280	4.4250±0.2375
EN	6.3114±0.2284	6.8372±0.2471	6.7859±0.2579	6.6288±0.2723	6.0472±0.3811	6.0361±0.4736	7.0678±0.3308
MI	2.8827±0.0889	3.1441±0.1525	2.8123±0.0360	2.6830±0.0976	2.1225±0.0213	2.3865±0.1383	4.8283±0.5393
SD	7.9249±0.2362	8.4949±0.2709	8.2895±0.2871	8.2699±0.2526	7.3311±0.5584	7.9917±0.2723	9.0244±0.4650
SF	0.0271±0.0006	0.0543±0.0120	0.0415±0.0018	0.0284±0.0012	0.0370±0.0020	0.024±0.0016	0.0516±0.0012
VIF	0.7037±0.0336	0.8345±0.0269	0.7914±0.0852	0.7595±0.0609	0.5608±0.0556	0.5325±0.0781	0.9622±0.0995
$N^{ab/f}$	0.007±0.005	0.02±0.008	0.007±0.004	0.002±0.003	0.002±0.001	0.01±0.004	0.015±0.006
SSIM	0.9±0.079	0.973±0.051	0.844±0.128	0.758±0.1	0.78±0.089	0.476±0.145	0.986±0.0037

according to different lighting conditions, retains the rich texture information and contrast information in the visible image while reducing the fusion of redundant information, and all these advantages are due to the design of the light-aware loss function and CMPF module.

2) QUANTITATIVE RESULTS

We select 50 image pairs from the M3FD dataset and 30 image pairs from the RoadScene dataset for quantitative comparison, respectively. The results of the quantitative comparison between our algorithm and the other six SOTA algorithms are shown in Figure 9 and Figure 10, respectively. In order to quantify the results of the comparison more clearly, we ranked these different algorithms according to eight evaluation metrics and showed the results in Table 2 and Table 3. From these results, it can be well seen that our algorithms achieve very good results in the four metrics AG, EN, SD, and SF, which indicates that our method has high fusion performance and strong generalization ability in different scenes. Our fused image contains not only the rich scene texture information in the visible image but also the contrast information of the saliency target in the infrared image; at the same time, among all the above algorithms, the fused image generated by our algorithm is the most consistent with the human visual photoreceptor system.

In conclusion, extensive qualitative and quantitative results on a variety of datasets show that our algorithm can not only adaptively select the effective information to fuse for different scenes, but it can also retain rich texture details and high contrast, providing the best visual quality in the resultant fused images.

F. INFRARED-VISIBLE OBJECT DETECTION

In this section, we will discuss the application of fused images in object detection tasks on the M3FD dataset. We use YOLOV7 [62] as a detector to detect the objects of interest in the image. At the same time, we also conducted quantitative and qualitative analyses of the detection results.

1) QUALITATIVE RESULTS

In Figure 11, under conditions of nighttime and smoke, the detector is unable to detect targets well from visible images, resulting in missed detections. On the contrary, under the same conditions, the thermal radiation information in infrared images is not affected, and sufficient target information is still retained. Therefore, the detector can detect the target, but the accuracy of the detected target is still not high. Among several fusion methods, from the perspective of visual effects, most of them effectively fuse the features of infrared and visible images without any missed detections. However, the SEDRFusion method has encountered false detections. Compared to the above methods, our method generates fusion images that are in line with the human visual system, achieving good detection results in various scenarios without any missed or false detections.

2) QUANTITATIVE RESULTS

The mean average precision and recall of different methods are quantitatively compared in Table 4. The proposed method achieved excellent detection accuracy on the M3FD dataset. Compared to infrared and visible images, there is a significant improvement in both recall and average detection accuracy. In terms of recall, it is 0.034 and 0.031 higher than the results in infrared and visible images, respectively. Regarding average accuracy, it is 0.052 and 0.046 higher than the results in infrared and visible images, respectively, and only 0.005 lower than the DenseFusion method.

G. EFFICIENCY COMPARISON

In Table 5, the runtimes of different fusion algorithms on the MSRS, M3FD, and RoadScene datasets are shown. All the methods were tested on the same platform, as given in Section IV-A. Among them, IFCNN, DenseFuse, and TarDAL are specially designed for real-time visual tasks, and they are all greatly optimized in terms of runtime speed, while our model is also a lightweight fusion network that can also achieve real-time image fusion. Compared to the other modeling algorithms, the images generated by our model are more consistent with the human visual system, and the

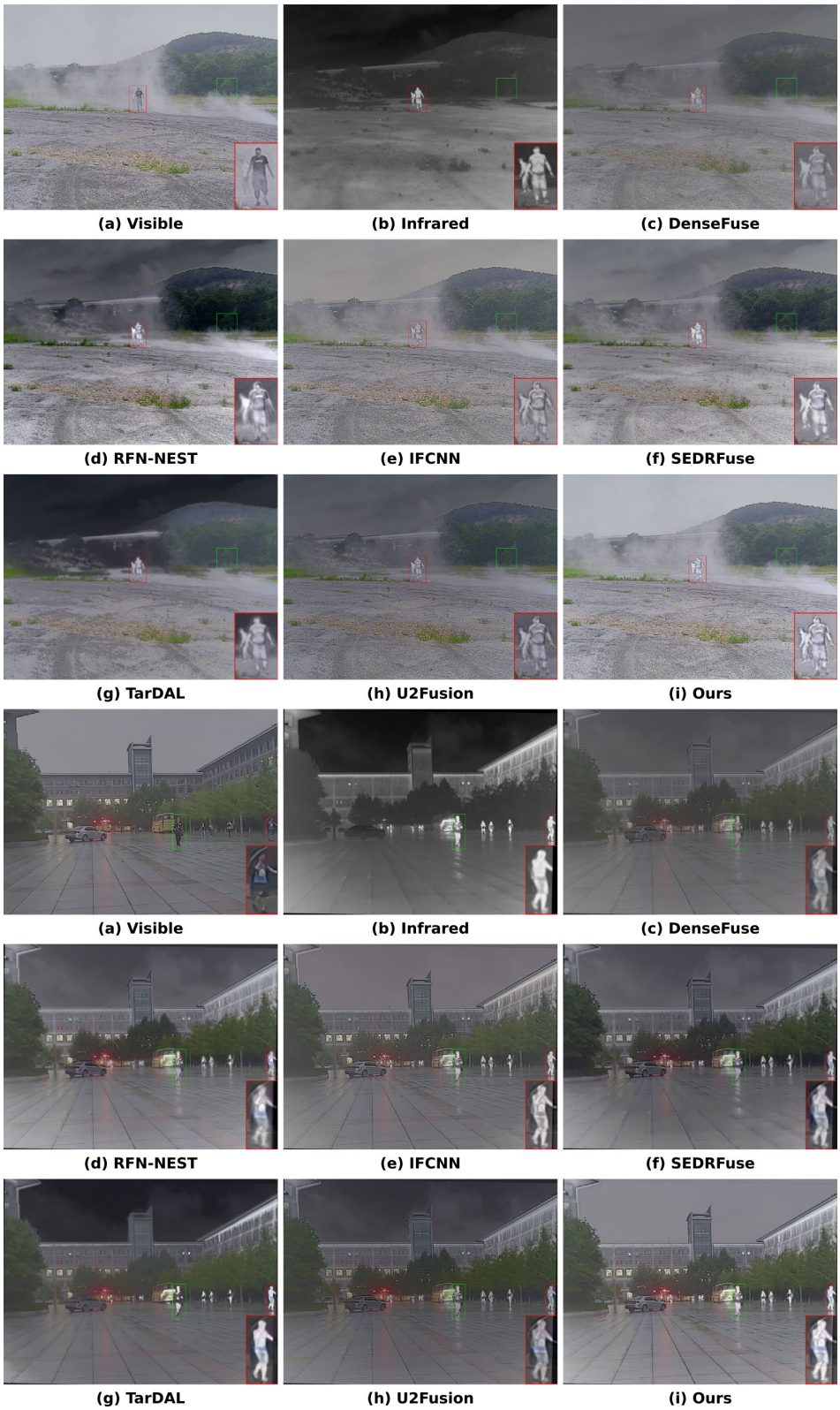


FIGURE 7. Qualitative comparison of GACnet with the other six SOTA methods on M3FD dataset. For visualization purposes, we select a detail region from each image and place its zoom-in at the bottom of the image (in a red box). We also select a saliency region from each image and highlight it in a green box.

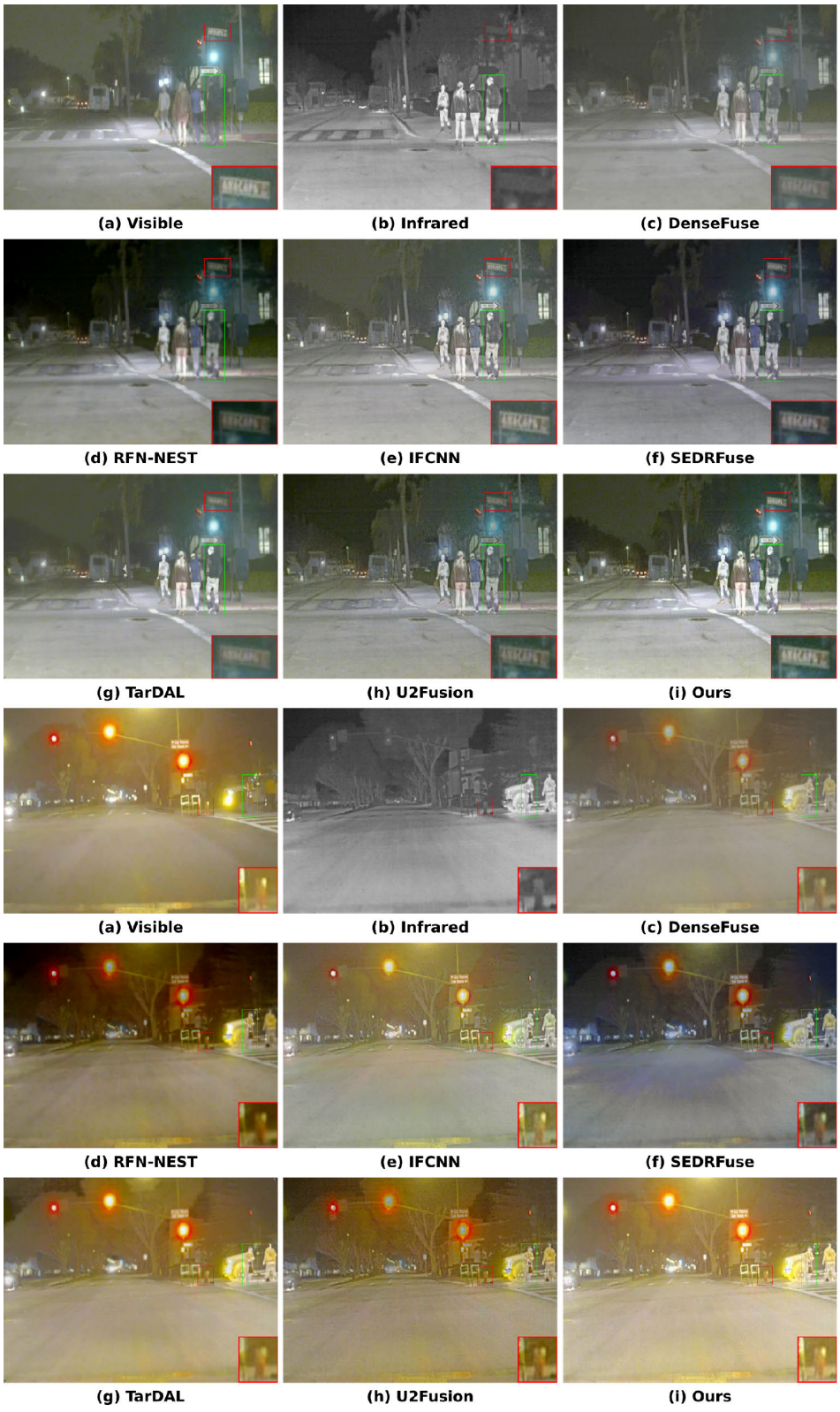


FIGURE 8. Qualitative comparison of GACnet on the RoadScene dataset with other SOTA six methods. For visualization purposes, we select a detail region from each image and place its zoom-in at the bottom of the image (in a red box). We also select a saliency region from each image and highlight it in a green box.

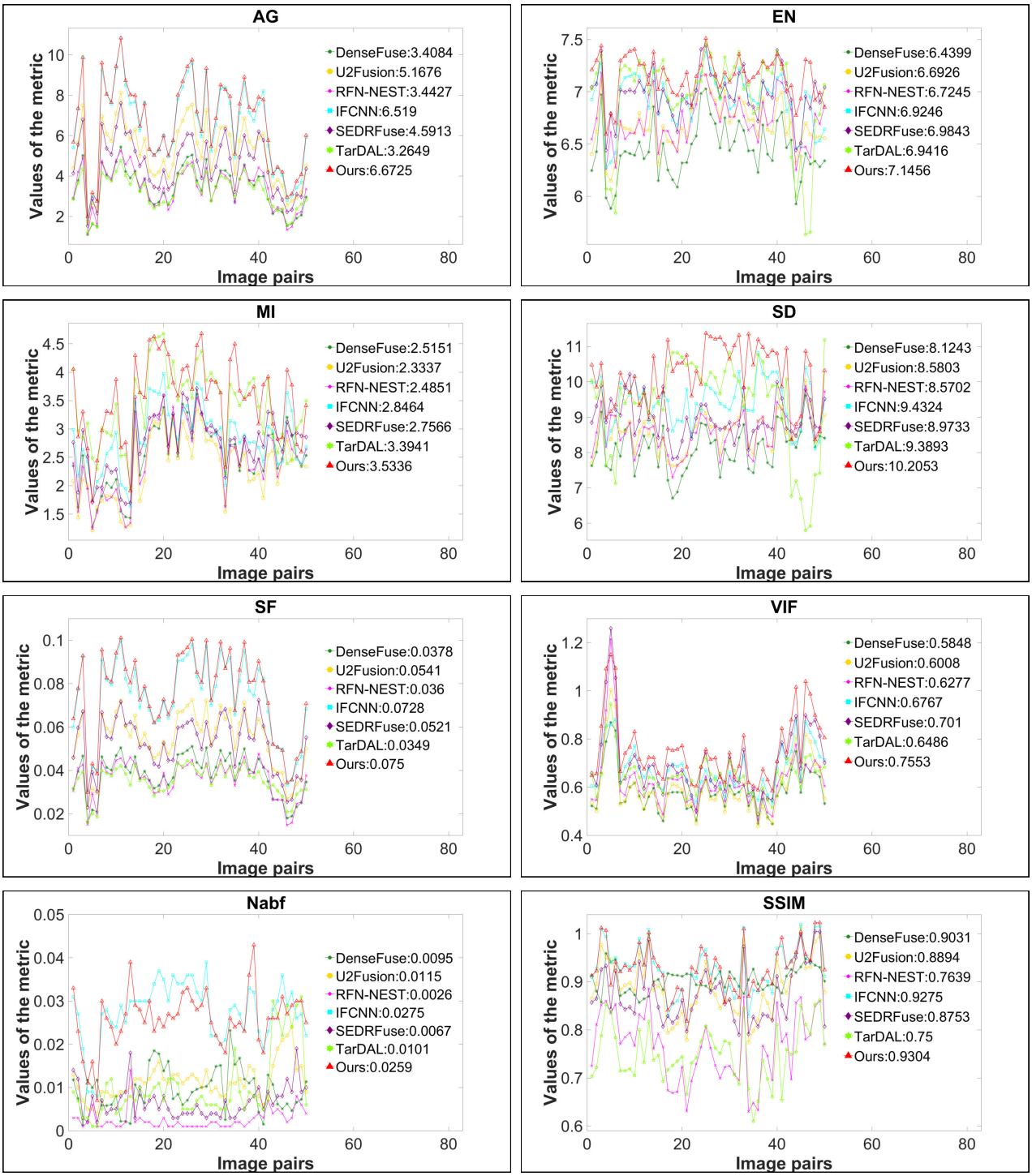


FIGURE 9. In order to clearly reveal the performance of the proposed approach and the six comparison methods, we only show the results of evaluating the eight metrics based on 50 pairs of images from the M3FD dataset. The higher the values of AG, EN, MI, SD, SF, VIF and SSIM, the lower the values of $N^{a/bf}$, the better the performance of the model algorithm.

performance metrics of our model are stronger in various aspects than those of the other six algorithms. In summary, a slight sacrifice in speed is acceptable.

H. ABLATION EXPERIMENT

In this section, we verify the effectiveness of the light loss function and the CMPF module.

TABLE 2. In order to clearly reveal the performance of the proposed approach and the six comparison methods, we only show the results of evaluating the eight metrics based on 50 pairs of images from the M3FD dataset. The higher the values of AG, EN, MI, SD, SF, VIF and SSIM, the lower the values of $N^{a/bf}$, the better the performance of the model algorithm.

Model	DenseFuse	IFCNN	SEDRFuse	RFN-NEST	U2Fusion	TarDAL	Ours
AG	6	2	4	5	3	7	1
EN	7	4	2	5	6	3	1
MI	5	3	4	6	7	2	1
SD	7	2	4	6	5	3	1
SF	5	2	4	6	3	7	1
VIF	7	3	2	5	6	4	1
$N^{ab/f}$	3	7	2	1	5	4	6
SSIM	3	2	5	6	4	7	1

TABLE 3. In order to clearly reveal the performance of the proposed approach and the six comparison methods, we only show the results of evaluating the eight metrics based on 30 pairs of images from the RoadScene dataset. The higher the values of AG, EN, MI, SD, SF, VIF and SSIM, the lower the values of $N^{a/bf}$, the better the performance of the model algorithm.

Model	DenseFuse	IFCNN	SEDRFuse	RFN-NEST	U2Fusion	TarDAL	Ours
AG	7	2	4	6	3	5	1
EN	7	5	3	2	6	4	1
MI	5	3	1	7	6	2	4
SD	7	6	2	3	4	5	1
SF	7	2	4	6	3	5	1
VIF	7	3	1	4	5	6	2
$N^{ab/f}$	1	5	4	2	6	3	7
SSIM	5	2	4	6	1	7	3

TABLE 4. A quantitative comparsion of different image fusion methods for object detection on the M3FD dataset.

Model	Visible	Infrared	DenseFuse	IFCNN	SEDRFuse	RFN-NEST	U2Fusion	TarDAL	Ours
Recall	0.719	0.716	0.779	0.769	0.771	0.753	0.771	0.748	0.75
mAP@.5	0.788	0.782	0.839	0.829	0.822	0.794	0.822	0.792	0.834
mAP@.5:.95	0.445	0.447	0.508	0.509	0.418	0.461	0.481	0.451	0.499

TABLE 5. Running time of different fusion algorithms on MSRS, M3FD, and RoadScene datasets. (in seconds @12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz and NVIDIA GeForce RTX 3060 Laptop GPU 7.8G).

Model	DenseFuse	IFCNN	SEDRFuse	RFN-NEST	U2Fusion	TarDAL	Ours
MSRS	0.2773	0.2574	6.8679	0.3786	1.2595	0.2533	0.3812
M3FD	0.4784	0.2598	15.2678	0.7222	2.8930	0.4332	0.5420
RoadScene	0.2808	0.1620	6.6144	0.4936	1.2679	0.3526	0.3826

TABLE 6. The changed light-aware loss function (after) and without changing the light-aware loss function (Before).

Model	AG	EN	MI	SD	SF	VIF	Nab/f	SSIM
After($\lambda_3 = 0.05$)	5.3049	7.1355	3.4093	11.8190	0.0500	0.7889	0.2688	0.9049
After($\lambda_3 = 0.07$)	5.1944	7.2349	3.5127	11.2124	0.0478	0.7632	0.2874	0.9193
Before($\lambda_3 = 0.01$)	5.4932	7.5397	3.7460	11.4718	0.0433	0.9700	0.2702	0.9523

1) ANALYSIS OF LIGHT LOSS FUNCTION

a: QUALITATIVE RESULTS

We designed a light-aware loss to guide the training of our fusion network. In order to verify the effectiveness of the light loss function, we conducted a series of ablation experiments and kept the other configurations unchanged and only changed the weight of the light loss function (λ_3) in the ablation experiments, and the results obtained are shown in Figure 12 and in Table 6. From the red and green boxes in Figure 12, we can find that the light loss function we designed

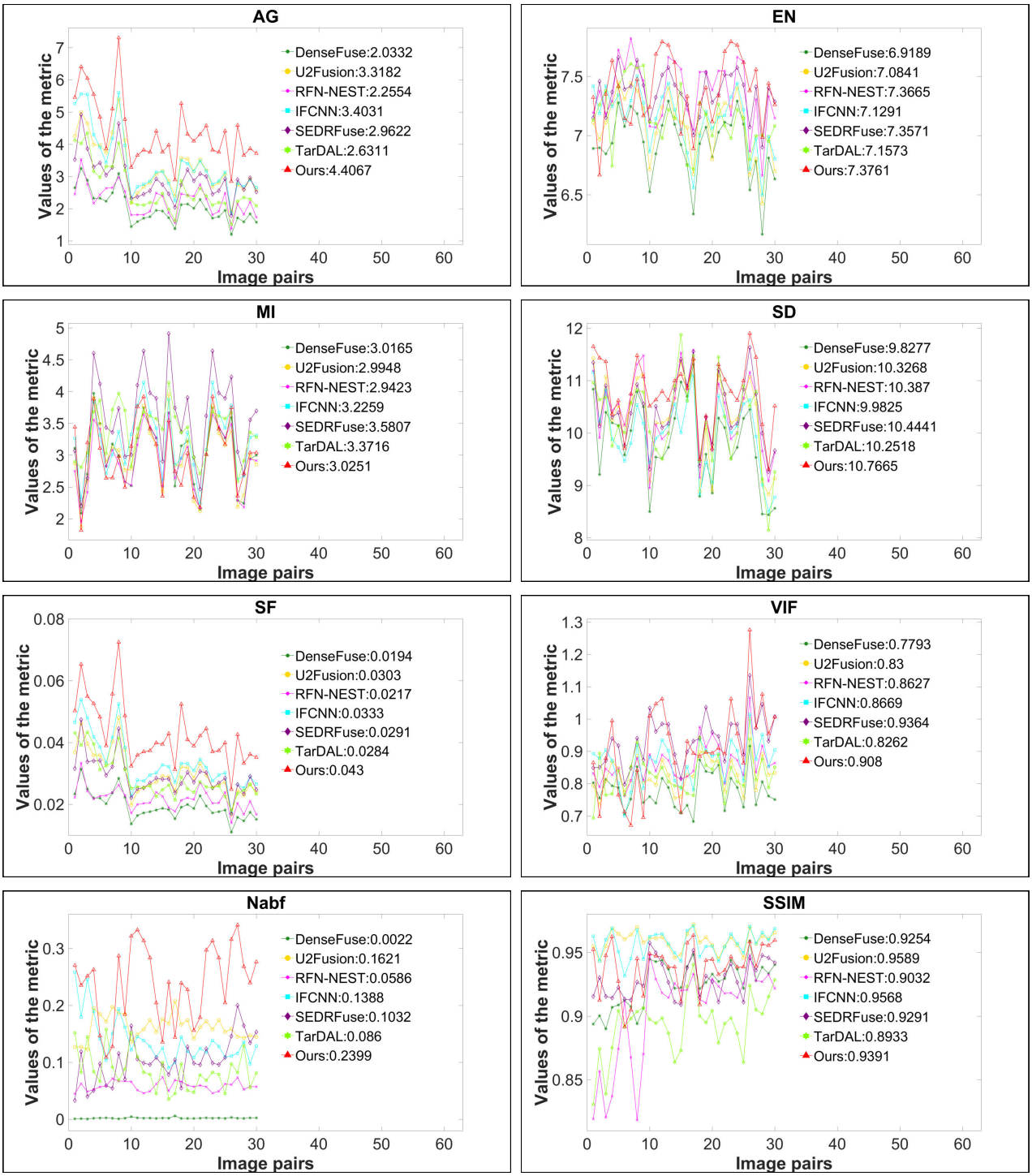


FIGURE 10. In order to clearly reveal the performance of the proposed approach and the six comparison methods, we only show the results of evaluating the eight metrics based on 30 pairs of images from the RoadScene dataset. The higher the values of AG, EN, MI, SD, SF, VIF and SSIM, the lower the values of $N^{a/bf}$, the better the performance of the model algorithm.

is indeed able to adaptively select effective information from infrared and visible images to guide the generation of fused images in different scenes according to different lighting conditions. We expect that the fusion image should retain more information about the visible image in a well-lit scene, because the visible image clearer details about the scene in this case. We would like the fused image to retain more information from the infrared image in poorly illuminated scenes, because the infrared image has more information about the scene in poorly illuminated scenes. As can be seen

FIGURE 11. A qualitative comparison of different image fusion methods for object detection on M3FD dataset.

from the red and green boxes in the Figure 12, after changing the λ_3 ($\lambda_3 = 0.05$ or 0.07), the fused image is greatly affected by the infrared radiation, and the branches of the tree and the ground become blurred, and the resulting image does not

FIGURE 12. The first column represents the infrared images, the second column represents the visible images, the third column represents the fusion images with the λ_3 changed ($\lambda_3 = 0.05$ and 0.07), and the fourth column represents the fusion image without changing the λ_3 ($\lambda_3 = 0.01$). The red and green boxes show the difference between the fusion results with the changed light-aware loss function (after) and without changing the light-aware loss function (Before).

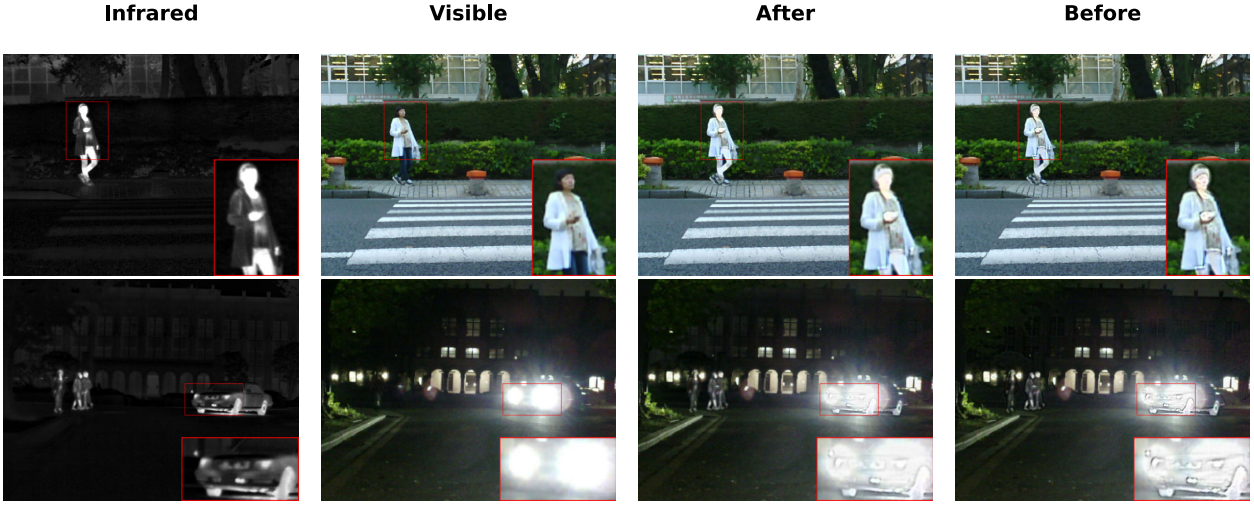


FIGURE 13. The first column represents the infrared image, the second column represents the visible images, the third column represents the fusion image after removing the CMPF module, and the fourth column represents the fusion image before removing the CMPF module. Also, the red box of the fused image shows the difference between the fused image obtained with and without the CMPF module. The red and green boxes show the difference between the fusion results without the CMPF module (after) and with the CMPF module (Before).

TABLE 7. The fusion results without the CMPF module and with the CMPF module.

Model	AG	EN	MI	SD	SF	VIF	Nab/f	SSIM
without CMPF	6.1364	7.4919	4.8045	9.6458	0.0667	0.8486	0.1592	0.9328
with CMPF	6.2567	7.5021	5.2512	9.6366	0.0685	0.9203	0.1565	0.9750

match our visual system. By contrast, with optimal λ_3 ($\lambda_3 = 0.01$), the fused image retains more target information that is salient in the infrared image in places with insufficient illumination. Where there is sufficient light, the fused image retains more detailed texture information in the visible image. This also shows that our designed light-aware network and loss function are very effective.

b: QUANTITATIVE RESULTS

The quantitative results of the light loss function ablation experiment are shown in Table 6. With λ_1 and λ_2 fixed, we tested the effect of the light loss function on the performance of the fusion network by changing the value of λ_3. From the quantitative analysis results corresponding to different λ_3 values, it can be seen that when $\lambda_3 = 0.01$ is

optimal. VIF measures the fidelity of the information from the point of view of human visual perception. This also meets our expectations.

2) CROSS MODAL INFORMATION FUSION MODULE ANALYSIS

a: QUALITATIVE RESULTS

In order to verify the effectiveness of the cross-modal information fusion module, we conducted a series of ablation experiments and kept other configurations unchanged during the ablation experiments. The results are shown in Figure 13. From the red and green boxes in Figure 13, we can find that our designed CMPF Module enhances the complementary information of the two modes, visible and infrared. From the red box in the first row, we can see that the fused image without removing the CMPF Module has more visible light texture information in the contours of the character's cheeks and clothes than the fused image after removing the CMPF Module. It is obvious from the red box in the second row that under the condition of severe light pollution in the visible image, the fused image without the CMPF module has more salient target information in the infrared image than the fused image with the CMPF module removed, and the outlines of the headlights are more obvious. This can be a strong indication that the CMPF module can enhance the fusion of complementary information between different modalities.

b: QUANTITATIVE RESULTS

The quantitative results of the CMPF module ablation experiment are shown in Table 7. It can be observed from the table that the fusion network with the CMPF module outperformers the fusion network without the CMPF module on various evaluation indicators. Among them, there is a significant improvement in the MI indicator.

V. CONCLUSION

In this study, we propose a Highly efficient spatial and radiation information mutual enhancing fusion method composed of three components: the adaptive light perception network (ALPnet), the gradient residual dense block (LGCnet), and the cross-modal perception fusion module (CMPF module). The ALPnet predicts the probability of light intensity in different scenarios, promoting the ability of the fusion network to generate high-quality fused images. The LGCnet enhances the ability to describe fine spatial details of fusion networks. The CMPF module promotes modal interaction in the fusion network and effectively enhances the fusion of common and complementary information between different modalities. Our method achieves a balance between fusion speed and fusion metrics. Extensive experiments have shown that our fusion method has competitive advantages over six SOTA deep-learning models in terms of visual effects and quantitative indicators.

REFERENCES

[1] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.
[2] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, and Y. Qiao, "Pedestrian detection with unsupervised multispectral feature learning using deep neural networks," *Inf. Fusion*, vol. 46, pp. 206–217, Mar. 2019.
[3] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 808–823.
[4] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, Jul. 2016.
[5] H. Li, X. Qi, and W. Xie, "Fast infrared and visible image fusion with structural decomposition," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106182.
[6] J. Ma and Y. Zhou, "Infrared and visible image fusion via gradientlet filter," *Comput. Vis. Image Understand.*, vols. 197–198, Aug. 2020, Art. no. 103016.
[7] Y. Liu, J. Jin, Q. Wang, Y. Shen, and X. Dong, "Region level based multi-focus image fusion using quaternion wavelet and normalized cut," *Signal Process.*, vol. 97, pp. 9–30, Apr. 2014.
[8] X. Liu, W. Mei, and H. Du, "Structure tensor and nonsubsampled shearlet transform based algorithm for CT and MRI image fusion," *Neurocomputing*, vol. 235, pp. 131–139, Apr. 2017.
[9] Q. Zhang and X. Maldague, "An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing," *Infr. Phys. Technol.*, vol. 74, pp. 11–20, Jan. 2016.
[10] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, Jan. 2020.
[11] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Inf. Fusion*, vol. 4, no. 4, pp. 259–280, Dec. 2003.
[12] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
[13] N. Yu, T. Qiu, F. Bi, and A. Wang, "Image features extraction and fusion based on joint sparse representation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1074–1082, Sep. 2011.
[14] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5792–5801.
[15] N. Cvejic, D. Bull, and N. Canagarajah, "Region-based multimodal image fusion using ICA bases," *IEEE Sensors J.*, vol. 7, no. 5, pp. 743–751, May 2007.
[16] J. Mou, W. Gao, and Z. Song, "Image fusion based on non-negative matrix factorization and infrared feature extraction," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, vol. 2, Dec. 2013, pp. 1046–1050.
[17] Z. Fu, X. Wang, J. Xu, N. Zhou, and Y. Zhao, "Infrared and visible images fusion based on RPCA and NSCT," *Infr. Phys. Technol.*, vol. 77, pp. 114–123, Jul. 2016.
[18] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 663–670.
[19] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[20] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.
[21] Q. Pan, L. Zhao, S. Chen, and X. Li, "Fusion of low-quality visible and infrared images based on multi-level latent low-rank representation joint with Retinex enhancement and multi-visual weight information," *IEEE Access*, vol. 10, pp. 2140–2153, 2022.
[22] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[23] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infr. Phys. Technol.*, vol. 82, pp. 8–17, May 2017.

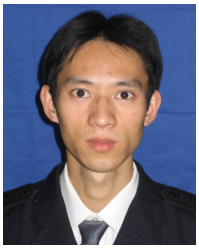
[24] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, May 2019.
[25] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.
[26] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.
[27] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[28] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
[29] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," 2020, *arXiv:2003.09210*.
[30] H. Wang, L. Li, C. Li, and X. Lu, "Infrared and visible image fusion based on autoencoder composed of CNN-transformer," *IEEE Access*, vol. 11, pp. 78956–78969, 2023.
[31] 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.
[32] 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, vol. 34, no. 7, pp. 12484–12491.
[33] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[34] H. T. Mustafa, P. Shamsolmoali, and I. H. Lee, "TGF: Multiscale transformer graph attention network for multi-sensor image fusion," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121789.
[35] D. Rao, T. Xu, and X.-J. Wu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Trans. Image Process.*, early access, 2023, doi: 10.1109/TIP.2023.3273451.
[36] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.
[37] 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.
[38] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.
[39] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
[40] H. Zhang, J. Yuan, X. Tian, and J. Ma, "GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1134–1147, 2021.
[41] X. Chen, Z. Teng, Y. Liu, J. Lu, L. Bai, and J. Han, "Infrared-visible image fusion based on semantic guidance and visual perception," *Entropy*, vol. 24, no. 10, p. 1327, Sep. 2022.
[42] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.
[43] 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.
[44] 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.
[45] J. Liu, Y. Wu, Z. Huang, R. Liu, and X. Fan, "SMoA: Searching a modality-oriented architecture for infrared and visible image fusion," *IEEE Signal Process. Lett.*, vol. 28, pp. 1818–1822, 2021.
[46] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.
[47] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.
[48] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.
[49] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–128, Dec. 1977.
[50] D. Sakkos, E. S. L. Ho, and H. P. H. Shum, "Illumination-aware multi-task GANs for foreground segmentation," *IEEE Access*, vol. 7, pp. 10976–10986, 2019.
[51] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 787–803.
[52] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[53] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.
[54] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.
[55] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[56] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, p. 355, 1997.
[57] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1102–1115, Apr. 2019.
[58] X. Zhang, P. Ye, and G. Xiao, "VIFB: A visible and infrared image fusion benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 468–478.
[59] 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, no. 4, pp. 600–612, Apr. 2004.
[60] B. K. S. Kumar, "Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform," *Signal, Image Video Process.*, vol. 7, no. 6, pp. 1125–1143, Nov. 2013.
[61] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, p. 313, 2002.
[62] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

ZONGZHEN LIU received the B.S. degree in electronic science and technology from Chengdu Technological University, Chengdu, China, in 2022. He is currently pursuing the M.S. degree with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu. His research interests include image fusion, object detection, and deep learning.

YUXING WEI is currently an Associate Professor with the Institute of Optics and Electronics, Chinese Academy of Sciences. His research interests include visual tracking, hardware/software systems, image processing, and understanding.

GELI HUANG received the B.S. degree in computer science and technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2022, where he is currently pursuing the M.S. degree with the School of Automation Engineering. His research interests include embedded AI accelerator architecture and image processing.

CHAO LI received the B.S. degree in automation engineering from the Hefei University of Technology, Hefei, China, in 2017. He is currently pursuing the M.S. degree with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include object detection and image fusion.



JIANLIN ZHANG received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Chengdu, China, in 2008. He is currently a Full Professor with the Institute of Optics and Electronics, Chinese Academy of Sciences. He has published more than 20 articles and conference papers in his research areas. His research interests include image processing and understanding, computer vision, machine learning, and artificial intelligence.

MEIHUI LI received the Ph.D. degree in signal and information processing from the University of Electronic and Science Technology of China (UESTC), Chengdu, Sichuan, China, in 2020. She is currently an Assistant Professor with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu. Her research interests include image processing, target detection, and tracking.

DONGXU LIU received the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2023. She is currently an Assistant Research Fellow with the Institute of Optics and Electronics, Chinese Academy of Sciences. Her research interests include deep learning and image classification.

XIAOMING PENG received the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, China, in 2005, and the Ph.D. degree from the University of Wollongong, Australia, in 2020. He is currently an Associate Professor with the School of Automation Engineering, University of Electronic Science and Technology of China. His research interests include image and video analysis, pattern recognition, computer vision, and deep learning.