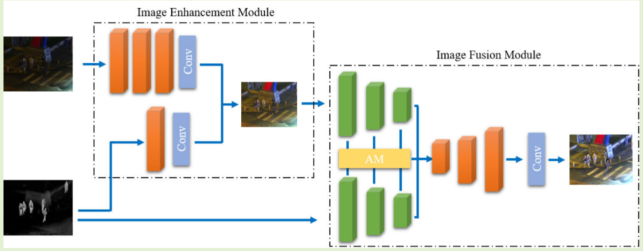
EV-Fusion: A Novel Infrared and Low-Light Color Visible Image Fusion Network Integrating Unsupervised Visible Image Enhancement

Xin Zhang, Xia Wang, Changda Yan, and Qiyang Sun



***Abstract*—Infrared and visible image fusion can effectively integrate the advantages of two source images, preserving significant target information and rich texture details. However, most existing fusion methods are only designed for well-illuminated scenes and tend to lose details when encountering low-light scenes because of the poor brightness of visible images. Some methods incorporate a light adjustment module, but they typically focus only on enhancing intensity information and neglect the enhancement of color feature, resulting in unsatisfactory visual effects in the fused images. To address this issue, this article proposes a novel method called EV-fusion, which explores the potential color and detail features in visible images and improve the visual perception of fused images. Specifically, an unsupervised image enhancement module is designed that effectively restores texture, structure, and color information in visible images by several non-reference loss functions. Then, an intensity image fusion module is devised to integrate the enhanced visible image and the infrared image. Moreover, to improve the infrared salient object feature in the fused images, we propose an infrared bilateral-guided salience map embedding into the fusion loss functions. Extensive experiments demonstrate that our method outperforms state-of-the-art (SOTA) infrared visible image fusion methods.**

***Index Terms*—Image fusion, infrared and visible image, nighttime environment, visible image enhancement.**

I. Introduction

THE image information obtained based on a single band or detector cannot represent all the information in the scene. Image fusion technology can achieve complementary advantages between different images, reduces data redundancy, and has great application prospects. Infrared and visible image fusion methods are currently a popular research direction with wide applications in military monitoring, traffic security, and other fields. Infrared images receive thermal radiation information from the scene, which is not affected by illumination and can highlight thermal target features in the scene. However, the image resolution is low with less texture information and lacks color information. Visible images contain rich texture and background information with higher resolution but less prominent target features. Moreover, under low illumination conditions, the visible images are dark with lower contrast. Therefore, the answer to the question of how to achieve infrared and visible image fusion in low-light environments is a key issue that this article addresses.

So far, a large number of methods for infrared and visible image fusion have been proposed, which can be roughly divided into two categories: traditional algorithms and deep learning-based algorithms. The core idea of traditional methods is to map the source image to a feature dimension using established image decomposition representation methods. Then, some feature fusion strategies are used for feature fusion before mapping the fused features back to an image [1]. Depending on the different image decomposition methods used, these algorithms can be further divided into multiscale transformation method [2], [3], sparse representation method [4], [5], subspace clustering method [6], optimization-based method [7], [8], and hybrid method [9]. However, these approaches have significant drawbacks as their fusion effect highly depends on manually designed feature extraction techniques that make it difficult to handle increasingly complex application scenarios. Additionally, their feature extraction approach is too singular with complex computing process leading to low operational efficiency.

The emergence of deep learning technology has recently advanced the development of image fusion techniques. Deep learning utilizes a data-driven approach to extract original fea-

Manuscript received 25 October 2023; accepted 13 December 2023. Date of publication 3 January 2024; date of current version 13 February 2024. The associate editor coordinating the review of this article and approving it for publication was Dr. Vinay Chakravarthi Gogineni. *(Corresponding author: Xia Wang.)*

The authors are with the Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing Institute of Technology, Beijing 100081, China (e-mail: zx1027533025@163.com; angelniuniu@bit.edu.cn).

Digital Object Identifier 10.1109/JSEN.2023.3346886

tures from source images and achieve better fusion effects [12]. The current mainstream methods for infrared and visible image fusion include autoencoder (AE)-based methods [13], [14], [15], end-to-end-based methods [10], [16], [17], [18], [19], [20], [21], [22], [23], and generative adversarial network (GAN)-based methods [24], [25], [26], [27], [28], [29], [30]. AE-based methods use encoders and decoders to extract and reconstruct image features while employing specific strategies for feature fusion. End-to-end-based methods achieve fusion effects by designing specific network structures and loss functions. GAN-based methods generate fused images using generators and adopt discriminators to constrain the fusion effectiveness. Moreover, to meet the application requirements in nighttime environments, enhancement-based methods [11], [31], [32], [33], [34] incorporate image enhancement modules to recover the image detail and contrast of the visible images.

While current deep learning-based methods can achieve relatively ideal image fusion effects, there are still some problems that need to be addressed. In this article, these issues are summarized into two main points.

1) First, the demand for infrared and visible image fusion scenes is mostly in low-light environments such as at night. Due to limited scene illumination, visible images often suffer from dark brightness, low contrast, missing texture details, and color distortion. Currently high-performance methods mainly focus on the daytime image fusion, which leads to severe degradation in fused images under low-light environment. Once the visible image is dark, its intensity will be much smaller compared to the infrared image. Therefore, the fused images cannot retain sufficient features of visible images, which is not the desired effect of image fusion.
2) The existing enhancement-based methods consider the recovery of visible images, but they only focus on the intensity feature and neglect the enhancement of color information. In the color image fusion task, we often extract the intensity component of the visible image and fuse it with the infrared image and the fused image exhibits significantly improved intensity contrast. However, if we maintain the original color information from visible image, the color contrast in the fused image will be much lower than the intensity contrast, resulting in a decrease in the quality of the fused image.

To address the challenging issues mentioned above, we propose a novel network for infrared and low-light color visible image fusion, named EV-fusion, which achieves joint training for color visible image enhancement and intensity image fusion. To achieve enhancement of both intensity and color information in the low-light visible image, we design an unsupervised image enhancement module. Specifically, we introduce a light-enhancement curve as our image enhancement model to enhance the visibility of visible images. This module takes dark visible images and infrared images as inputs, producing a light-enhancement map as output. To ensure effective image enhancement, we design specific non-reference loss functions based on intensity and color. Next, we input the enhanced visible intensity component and infrared image into the image fusion module, which is based on the swin transformer architecture, ensuring excellent feature extraction performance. In addition, to explore the salient feature of the infrared images, a bilateral-guided salience map is introduced in fusion loss function by adopting bilateral-guided filtering for extracting infrared target regions. Finally, the fused image will be obtained by combining fused intensity components with enhanced color components. To validate the effectiveness of the proposed joint training model in this article, we not only compare it with other image fusion methods but also conduct a comparison with two-stage fusion strategy that combines mainstream low-light image enhancement methods with existing image fusion methods. From the analysis of all experimental results, it can be observed that our EV-fusion effectively incorporates the differences between enhancement and fusion modules.

Overall, the main contributions of this article are as follows.

1) We propose an infrared and low-light color visible image fusion method suitable for nighttime environments, which achieves joint training of color visible image enhancement and intensity image fusion, obtaining high-contrast color image fusion in low-light scenarios.
2) We propose an unsupervised color image enhancement module that takes infrared and low-light visible images as inputs to improve both intensity and color information in visible images.
3) In order to improve the salient object features of fused images, we design a salience map extraction method based on bilateral-guided filtering to guide the image fusion.
4) The experiment shows that the fusion performance of this article has better brightness, contrast, color, and naturalness compared to other state-of-the-art (SOTA) methods. Furthermore, the comparison with two-stage fusion strategy demonstrates that our method is able to effectively integrate the differences between enhancement and fusion modules.

II. Related Work

In this section, we mainly review the existing infrared and visible image fusion methods.

A. Deep Learning-Based Image Fusion Methods

Currently, high-performance image fusion methods are mostly based on deep learning, which can be divided into three main categories, including AE-based methods, end-to-end-based methods, and GAN-based methods. Densefuse [13] was the first attempt at utilizing AEs, incorporating dense modules in the encoder to extract more useful information from source images. RFN-nest [15] uses residual fusion networks instead of previously manually designed feature fusion strategies, along with detail preservation loss functions and feature enhancement loss functions to achieve this goal. IFCNN [19] proposes a unified model based on CNN structure to implement multitask image fusion algorithm. U2fusion [10] can adaptively estimate the importance of different source images and generate corresponding feature weights. SEAfusion [17] cascades the image fusion module with semantic segmentation

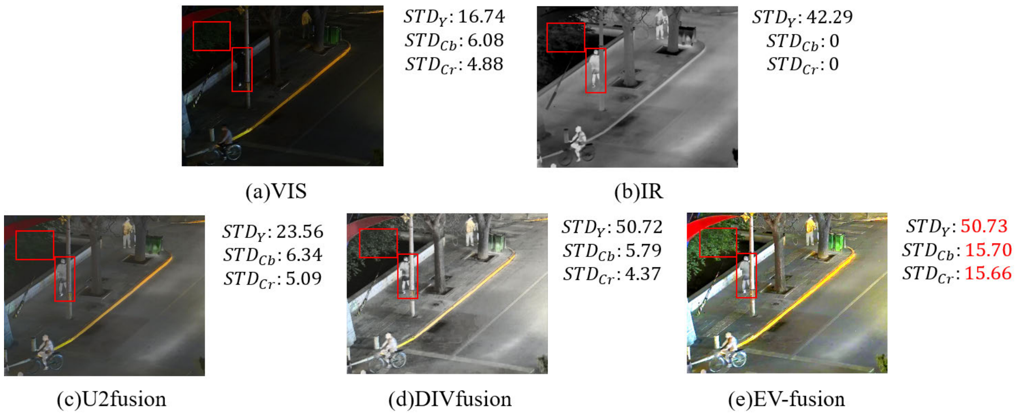


Fig. 1. Infrared and low-light color visible image fusion results. U2fusion [10] is an end-to-end-based method without visible image enhancement, DIVfusion [11] is an enhancement-based method. The top-right corner of each plot displays the contrast of the three channels (Y, Cb, and Cr) in YCbCr color space by STD, where STD$_Y$ represents the intensity contrast, STD$_{Cb}$ and STD$_{Cr}$ represent the color contrast. It can be observed that the contrast of the fused image by U2fusion is only slightly enhanced compared to the visible image. Although DIVfusion can effectively enhance the intensity contrast, the contrast of the two color components is even reduced. Our proposed method, on the other hand, can effectively enhance both intensity and color features. (a) VIS. (b) IR. (c) U2fusion. (d) DIVfusion. (e) EV-fusion.

module, using semantic loss to guide high-level semantic information backflow to the image fusion module, effectively improving the performance of advanced visual tasks on fused images. SwinFusion [21] proposed a new general image fusion framework based on cross-domain long-range learning and swin transformer. FusionGAN [24] establishes an adversarial game between the generator and discriminator, where the goal of the generator is to generate fused images with primary infrared intensity and additional visible gradients, while the goal of discriminator is to force more details from visible images into fused images. These methods provide different approaches for fusion strategies. In general, end-to-end-based methods can better fulfill specific requirements compared to the other two methods by designing specific forms of loss functions.

B. Enhancement-Based Image Fusion Methods

Enhancement-based methods aim at recovering the lost information from the dark visible images. GFCE [31] presents a night-vision context enhancement algorithm with the guided filter. HMD-ALA [34] combines both local and global contrast enhancements and introduces an adaptive light adjustment algorithm. VDFEFuse [35] creates a visual differentiation feature extraction operator to compensate for the loss of important features. PIAfusion [32] designs an illumination-aware subnetwork to estimate illumination distribution and calculate illumination probability, thereby adaptively maintaining intensity distribution of salient targets while preserving texture information in background areas. DIVfusion [11] combines image enhancement module with image fusion for the first time, achieving enhancement in dark areas and restoring color features of images.

In summary, although current image fusion methods can achieve good complementary advantages between infrared and visible images, they lack consideration for degradation of color visible images under low-light conditions. This seriously affects the quality and visibility of the fused images. To the best of our knowledge, only DIVfusion [11] realizes color image fusion, but it does not enhance color features but rather preserves initial colors only.

III. Method

In this section, we describe our EV-fusion in detail.

A. Motivation

To better illustrate the aforementioned problem, we provide an example to validate our viewpoint that enhancement of intensity and color is necessary for infrared and low-light color visible image fusion. As shown in Fig. 1, we compare the image contrast of different fusion results. Here, standard deviation (STD) is a metric for measuring the contrast of an image, which represents the degree of deviation between the pixel values of the entire image and their mean value. To distinguish intensity and color information, we converted the RGB image to the YCbCr color space for comparison, where STD$_Y$ represents the intensity contrast, STD$_{Cb}$ and STD$_{Cr}$ represent the color contrast. It can be seen that existing fusion results have not achieved significant contrast enhancement from both subjective and objective perspectives. Non-enhancement method U2fusion can only obtain low-contrast fused images. Although DIVfusion enhances the intensity of the visible image, it only preserves the original visible color components, resulting in a decrease in color contrast in the fused image.

To meet the requirements of fusing infrared and color visible images in low-light environments, we have to consider the enhancement of intensity and color information simultaneously. To this end, we propose a joint model including two modules: a color visible image enhancement module and an intensity image fusion module. The two modules are jointly trained to facilitate the sharing of underlying feature representations between two tasks. By sharing features, the knowledge and representations learned from one task can be transferred to the other task, thereby enhancing the performance of both tasks. Fig. 2 illustrates the overall flowchart of the network. The core idea of the visible image enhancement is to improve intensity and color features in low-light images. It takes both

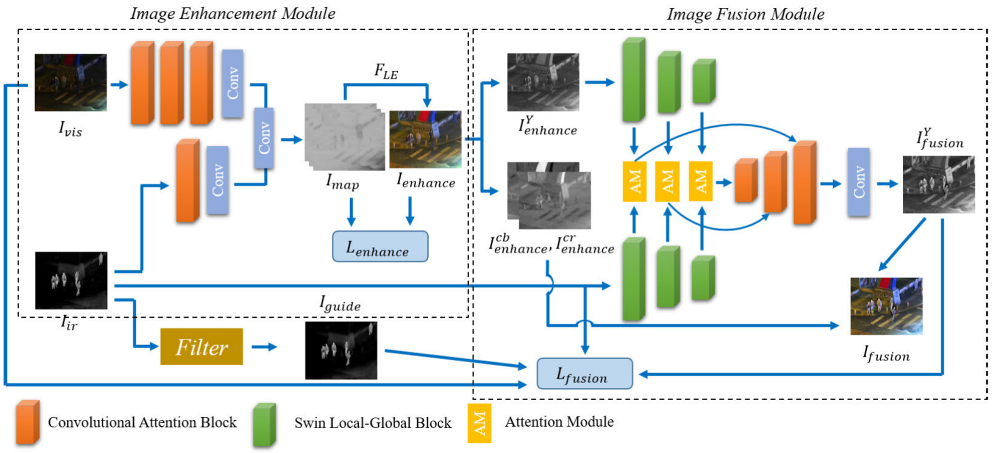


Fig. 2. Overall structure of EV-fusion.

infrared and visible images as inputs, extracts features from the source images through their respective independent sub-networks, finally fuses the output of the sub-networks to generate a light-enhancement map. This map is then used to enhance the visible image by the light-enhancement curve [36]. Afterward, color conversion is performed on this image by decomposing it into Y component and CbCr component. The CbCr components are retained while the decomposed Y component and infrared image are inputted together into the image fusion module. In image fusion module, multimodal features from different source images need to be integrated. So, multiscale specific feature extraction modules are designed to extract corresponding multiscale feature maps from different source images, which are integrated at different scales using attention modules (AMs). The fused intensity images will be reconstructed by feature reconstruction module. Finally, the preserved color components after enhancement are combined with fused intensity component to obtain the final colored fusion image.

B. Color Visible Image Enhancement Module

In recent years, numerous studies have focused on low-light image enhancement [36], [37], [38], [39], [40]. These algorithms can efficiently restore brightness, texture, and color in low-light images. In this article, we propose an unsupervised image enhancement module that utilizes both infrared and visible images embedded before the fusion modules. Inspired by Guo et al. [36], our method employs a light-enhancement curve as the fundamental model for visible image enhancement. This curve can transfer the dark visible image to its enhanced version automatically, where the light-enhancement map is solely dependent on the input. The light-enhancement curve can be expressed as

$$F_{\mathrm{LE}}\left(I_{\mathrm{vis}}\right) = I_{\mathrm{vis}} + I_{\mathrm{map}} * I_{\mathrm{vis}} * \left(1 - I_{\mathrm{vis}}\right) \tag{1}$$

where I_{vis} denotes the dark visible image, I_{map} represents the pixel-wise light-enhancement map. $F_{\mathrm{LE}}()$ means the enhancement curve process. This curve is monotonous to preserve the contrast of the original image. To ensure the natural color perception of fused images, this article sets the I_{map} in a three-channel format.

Given a visible image I_{vis} and an infrared image I_{ir}, first, the two images are, respectively, input into different sub-networks. The sub-network for visible images contains three convolutional attention modules (CABs) and one convolutional layer, while the sub-network for infrared images contains one CAB and one convolutional layer. The schematic of CAB is shown in Fig. 3. Motivated by modern low-level vision tasks [41], we add spatial and channel AMs based on Res-block [42], which share information within a feature tensor in terms of both spatial and channel dimensions. The CAB is able to extract informative local features and suppress redundant ones. The spatial/channel AM aims to generate a spatial/channel attention map by average and global pooling to rescale the input feature map. The outputs of the sub-networks are merged by a convolutional layer to obtain the pixel-wise light-enhancement map I_{map}, which will be used to enhance the image by applying the light-enhancement curve. The whole process of visible image enhancement can be expressed as

$$\begin{cases} I_{\mathrm{map}} = \mathrm{Conv}\left(\psi_{\mathrm{vis}}\left(I_{\mathrm{vis}}\right), \psi_{\mathrm{ir}}\left(I_{\mathrm{ir}}\right)\right) \\ I_{\mathrm{enhance}} = F_{\mathrm{LE}}\left(I_{\mathrm{vis}}, I_{\mathrm{map}}\right) \end{cases} \tag{2}$$

where $\psi_{\mathrm{vis}}()$ represents the visible sub-network while $\psi_{\mathrm{ir}}()$ represents the infrared sub-network. I_{enhance} denotes the enhanced visible image.

C. Intensity Image Fusion Module

First, the enhanced visible image I_{enhance} obtained by former module is decomposed into Y component $I^{\mathrm{Y}}_{\mathrm{enhance}}$, Cb component $I^{\mathrm{Cb}}_{\mathrm{enhance}}$, and Cr component $I^{\mathrm{Cr}}_{\mathrm{enhance}}$ through color transformation. The input of the image fusion module is the enhanced Y component and infrared image, which are, respectively, subjected to swin local-global block (SLGB). Due to the different computational mechanisms of CNN and transformer, CNN tends to extract local information from images while transformer tends to extract global information [43], [44]. In order to maximize the effective information obtained from the source images, we design SLGB. Moreover, to expand the receptive field of feature extraction, we introduce

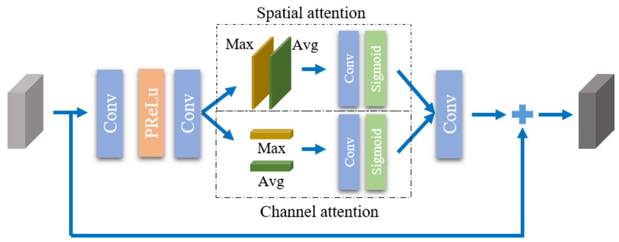


Fig. 3. Structure of CAB.

a multiscale structure in feature extractors, which can be formulated as

$$\begin{cases}\phi_{\text{vis}} = \Phi_{\text{SLG}}^{\text{vis}}\left(I_{\text{enhance}}^{\text{Y}}\right) \\ \phi_{\text{ir}} = \Phi_{\text{SLG}}^{\text{ir}}\left(I_{\text{ir}}\right)\end{cases} \tag{3}$$

where Φ_{vis} and Φ_{ir} represent the visible feature extractor and infrared feature extractor, respectively. Then, we integrate the output features ϕ_{vis} and ϕ_{ir} from different scales through AMs (AE) for feature reconstruction. Finally, the fused intensity image is obtained. The above process can be expressed as

$$\begin{cases}\phi_{\text{fuse}} = \Phi_{\text{AE}}\left(\phi_{\text{vis}}, \phi_{\text{ir}}\right) \\ I_{\text{fuse}}^{\text{Y}} = \Phi_{\text{FR}}\left(\phi_{\text{fuse}}\right)\end{cases} \tag{4}$$

where Φ_{AE} represents the AM and Φ_{FR} represents the feature reconstruction module. $I_{\text{fuse}}^{\text{Y}}$ denotes the fused intensity image, which will be combined with the retained Cb component $I_{\text{enhance}}^{\text{Cb}}$ and Cr component $I_{\text{enhance}}^{\text{Cr}}$ to obtain the final fused color image I_{fuse}

$$I_{\text{fuse}} = \text{YCbCr2RGB}\left(I_{\text{fuse}}^{\text{Y}}, I_{\text{enhance}}^{\text{Cb}}, I_{\text{enhance}}^{\text{Cr}}\right). \tag{5}$$

1) Swin Local–Global Feature Extractor: Transformer is a popular network model that has been applied in the field of infrared and visible image fusion. In order to integrate both local and long-range dependency features from different source images, we design the SLGB inspired by swin transformer [45]. Specifically, three SLGBs of different scales are employed in each feature extractor, which will output three outputs of different scales to maintain multiscale features. A down-sampling operator is embedded during the patch-merging. The structure of SLGB is illustrated in Fig. 4. First, we introduce a parallel convolution block for local features, which contains convolutional layers with different kernel sizes, namely 3 × 3, 5 × 5, and 7 × 7. Denoting the input feature map as x_{in}, the parallel convolution block can be formulated as

$$\begin{aligned} x_{l1} &= \text{Conv}_{3\times 3}\left(x_{\text{in}}\right) \\ x_{l2} &= \text{Conv}_{5\times 5}\left(x_{\text{in}}\right) \\ x_{l3} &= \text{Conv}_{7\times 7}\left(x_{\text{in}}\right) \\ x_{\text{local}} &= \text{Conv}_{1\times 1}\left(\text{concat}\left(x_{l1}, x_{l2}, x_{l3}\right)\right). \end{aligned} \tag{6}$$

Then, the features will be input into a multihead self-attention block, which is employed to extract global features. Taking the feature map $x_{\text{conv}} \in \text{R}^{H\times W\times C}$ from the local convolutional block, we reshape it by patch-merging it into non-overlapping local windows of size $M \times M$. This will create a new tensor with dimensions $(HW)/(M^2) \times M^2 \times C$, where we set $M = 8$. The self-attention mechanism is computed separately for each window. In the case of window feature $x_i \in \text{R}^{M^2\times C}$ where $i = 1, \ldots, N$, query Q_i, key K_i, and value V_i are defined as follows:

$$Q_i = x_i F_Q, \quad K_i = x_i F_K, \quad V_i = x_i F_V \tag{7}$$

where $F_Q, F_K, F_V \in \text{R}^{C\times d}$ are projection matrices for all the local windows. From these, the traditional self-attention matrix is computed by

$$\text{SA}\left(Q_i, K_i, V_i\right) = \text{softmax}\left(\frac{Q_i \cdot K_i^T}{\sqrt{d}} + \text{B}\right) \cdot V_i. \tag{8}$$

In this context, B is a position coding parameter that can be learned, while d represents the dimension of the key feature. Following the multihead self-attention mechanism, we parallelly apply self-attention h number of times to generate distinct attention distributions. Our work utilizes $h = 3$. Before and after both MSA and feed forward network (FFN), we incorporate layernorm (LN) and residual skipping connection. Finally, we conduct feature extraction with an FFN composed of two fully connected layers and Gaussian error linear units (GELUs). The entire pipeline of window multihead self-attention (WMSA) is expressed as

$$\begin{aligned} Z_i &= \text{WMSA}\left(\text{LN}\left(Q_i, K_i, V_i\right)\right) + x_i \\ Z_i &= \text{FFN}\left(\text{LN}\left(Z_i\right)\right) + Z_i. \end{aligned} \tag{9}$$

The existing WMSA mechanism limits the self-attention to each individual window, which overlooks relevant information across different windows. To address this issue, we introduce a modified version called the shifted WMSA module (SWMSA) [46], which shifts the window location by $([(M)/(2)], [(M)/(2)])$ pixels during partitioning. This approach promotes information exchange between windows by incorporating shifted window positions. SWMSA is similar to WMSA in terms of formulation

$$\begin{aligned} Z_i &= \text{SWMSA}\left(\text{LN}\left(Q_i^Z, K_i^Z, V_i^Z\right)\right) + Z_i \\ Z_i &= \text{FFN}\left(\text{LN}\left(Z_i\right)\right) + Z_i. \end{aligned} \tag{10}$$

Moreover, the self-attention mechanism of SLGB has a limited receptive field due to window shifting within two fixed windows. However, multiscale learning can extend the receptive field scale. To maintain consistency between global and local features, parallel convolution modules were also introduced before SWSA [47].

2) Attention Module: After obtaining heterogeneous features from different dimensions, we use AMs to aggregate features of different scales separately. The structure diagram of the AM is shown in Fig. 5. After directly adding two features, spatial and channel weights are obtained through spatial attention and channel attention, respectively. These two weights are assigned to the initial feature map, which is further processed through another convolutional layer to produce an aggregated feature output. The above process can be expressed as

$$\begin{aligned} \phi_{\text{sum}} &= \phi_{\text{vis}} + \phi_{\text{ir}} \\ \phi_{\text{spa}} &= \Phi_{\text{spa}}\left(\phi_{\text{sum}}\right) \end{aligned}$$

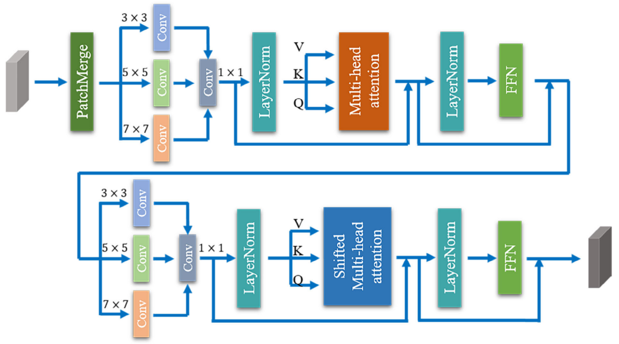


Fig. 4. Structure of SLGB.

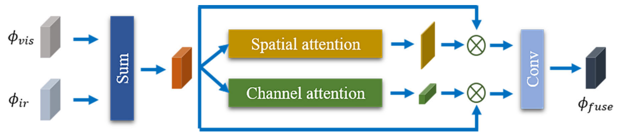


Fig. 5. Structure of AM.

$$\begin{aligned}\phi_{\text{cha}} &= \Phi_{\text{cha}}\left(\phi_{\text{sum}}\right)\\ \phi_{\text{fuse}} &= \text{Conv}\left(\text{Concat}\left(\phi_{\text{spa}} * \phi_{\text{sum}}, \phi_{\text{cha}} * \phi_{\text{sum}}\right)\right). \qquad (11)\end{aligned}$$

3) Feature Reconstruction Module: After obtaining the features aggregated at different scales, we use a multiscale feature reconstruction module to generate the fused image. Specifically, we sequentially input three different scale features into the CAB, and introduce a patch-folding operator between every two modules to transform feature scales, which is the inverse operation of the patch-merging in SLGB. Finally, we use one convolutional layer to reconstruct the fused image.

D. Loss Functions

Since both visible image enhancement and image fusion do not have labels and follow the rules of unsupervised training, the design of the loss function plays a decisive role in the effectiveness of the algorithm. As described before, we integrate the loss functions of both modules to achieve joint training. This section introduces the loss functions, respectively.

1) Loss Functions for Visible Image Enhancement Module: To better recover brightness, detail, and color characteristics of visible images and ensure the effectiveness of the visible image enhancement module, we propose four loss functions to measure the quality of the output image.

a) Illumination loss: To restore the brightness of underexposed areas in the image, we designed an illumination loss [48]. This loss function controls the brightness of the image by controlling the average pixel value within a local area of the grayscale image, which can be expressed as

$$L_{\text{illu}} = \frac{1}{N}\sum_{i}^{N}\left|I_i - C\right| \qquad (12)$$

where N represents the number of local areas, whose size is 16×16. I_i represents the average pixel value of each local area. C denotes the ideal illumination value of each area, which we set to 0.5 in our paper.

b) Color loss: To ensure the natural color appearance of the enhanced image, we introduce a color loss. According to the Gray-World color constancy hypothesis that the color in each sensor channel is averaged to gray over the entire image [49], we introduce a color loss to limit the correlation between different color channels, in order to adjust the color of the enhanced image. The color loss can be formulated as

$$L_{\text{col}} = \sum_{\forall(m,n)\in\varepsilon}\left(J^m - J^n\right)^2, \quad \varepsilon = \{(\text{R}, \text{G}), \ (\text{R}, \text{B}), (\text{G}, \text{B})\} \qquad (13)$$

where J^m represents the average value of m channel of enhanced visible image.

c) Smoothness loss: In order to maintain spatial consistency within the enhanced image and prevent noise in the image from being amplified, we introduce a smoothness loss based on light-enhancement map [36], which can be formulated as

$$L_{\text{tv}} = \frac{\sum\left|\nabla_h I_{\text{map}}\right|^2}{(H-1)*W} + \frac{\sum\left|\nabla_w I_{\text{map}}\right|^2}{H*(W-1)} \qquad (14)$$

where H and W denote the horizontal and vertical pixels in light-enhancement map, ∇_x and ∇_y represent the horizontal and vertical gradient operations, respectively.

2) Bilateral-Guided Salience Map for Infrared Images: Before introducing the loss functions of the fusion module, we need to emphasize the proposed bilateral-guided salience map. It can extract salient regions with clear edges from the infrared image, helping the fusion module to focus more on the salient features of the infrared image. This salience map employs a combination of bilateral filter and guided filter [50] to measure prominent target areas in infrared images. In this article, we embed this bilateral-guided salience map into the loss functions as the objective of network optimization, rather than using it as a pre-processing module as in previous algorithms. On one hand, this reduces computational steps and improves overall algorithm efficiency. On the other hand, we still retain the original features of the infrared image in the fusion module, providing the algorithm with more flexibility and manipulation capabilities. This map can be expressed as

$$I_{\text{guide}} = F_{\text{guide}}\left(I_{\text{ir}}, F_{\text{bila}}\left(I_{\text{ir}}\right)\right) \qquad (15)$$

where F_{guide} means the guided filter, while F_{bila} means the bilateral filter. Both of these filters are mainstream edge-preserving filters. As indicated in Fig. 6, we adopt the result of bilateral filter as the guide image in guided filter, which is more effective than using only itself as the guide image. Subsequently, we normalize the salience map to [0, 1] as weights for visible images, thereby limiting the proportion of visible images in the target area.

3) Loss Functions for Intensity Image Fusion Module: In order to meet the feature requirements of different source images for fused images, we propose two loss functions to measure the correlation between fused images and source images.

a) Structure loss: To restore the structural features of the original image, we employ multiscale structural similarity (MS-SSIM) to measure the image structure of the fusion

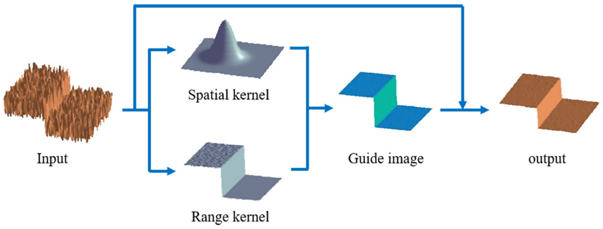


Fig. 6. Pipeline of the bilateral-guided salience map.

results. Here, we use five different scales, which are 3, 5, 7, 9, and 11, respectively. In previous image fusion methods, the weights of the loss functions between the fused image and different source images were the same. However, this could result in the fused image appearing more similar to an average of the two images, or even cause problems such as information loss due to overexposure of the visible image. To balance the characteristics of the source images, we add the salience map I_{guide} to the visible parts of the loss function

$$L_{\text{str}} = 1 - \text{msssim}\left(I^{\text{Y}}_{\text{fusion}}, I_{\text{ir}}\right) + \left(1 - \text{msssim}\left(I^{\text{Y}}_{\text{fusion}}, I^{\text{Y}}_{\text{enhance}}\right)\right) * \left(1 - I_{\text{guide}}\right) \quad (16)$$

where I_{guide} represents the pixel-wise salience map of infrared image.

b) Gradient loss: One of the main purposes of image fusion is to obtain more texture and detail information from the source images, so we introduce gradient loss, which is formulated as

$$L_{\text{grad}} = \|\nabla I^{\text{Y}}_{\text{fusion}} - \max\left(|\nabla I^{\text{Y}}_{\text{enhance}}|, |\nabla I_{\text{ir}}|\right)\|_1 \quad (17)$$

where ∇ denotes the Laplacian operator, max() refers to the element-wise maximum selection.

4) Total Loss Function: Our method jointly trains the visible image enhancement module and image fusion module, so the total loss is the weighted sum of all sub-losses mentioned before, which is expressed as

$$\begin{cases} L_{\text{enhance}} = \lambda_{\text{illu}} L_{\text{illu}} + \lambda_{\text{col}} L_{\text{col}} + \lambda_{\text{tv}} L_{\text{tv}} \\ L_{\text{fusion}} = \lambda_{\text{str}} L_{\text{str}} + \lambda_{\text{grad}} L_{\text{grad}} \\ L_{\text{total}} = L_{\text{enhance}} + L_{\text{fusion}} \end{cases} \quad (18)$$

where λ_t denotes the coefficients corresponding to different sub-loss functions. Empirically, we set $\lambda_{\text{illu}} = 10, \lambda_{\text{col}} = 6, \lambda_{\text{tv}} = 200, \lambda_{\text{str}} = 1$, and $\lambda_{\text{grad}} = 1$.

IV. Experiments and Results

A. Datasets and Implementation Details

We adopt LLVIP [51] and MSRS [32] as the dataset for our paper. Among them, LLVIP is a paired infrared and color visible image dataset under low-light environments, including 16 836 pairs of images. We randomly selected 10 000 pairs of images as the training set for our fusion model, with visible images as the training set for our enhancement model. Then, we randomly selected 150 pairs of images from the remaining images as our first validation set. The MSRS dataset contains 715 pairs of daytime images and 729 pairs of nighttime images. We directly used the 180 paired nighttime images in the test set as our second validation set. Therefore, the validation set here consists of two parts with a total of 320 image pairs.

During training process, we use the Adam optimization method to optimize the parameters, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial value of the learning rate is 1×10^{-4}. In order to better acquire global and semantic features and reduce edge errors caused by image cropping, all images are resized to 256 × 256. All experiments are performed using two NVIDIA GeForce RTX 3090 GPUs.

B. Comparison of SOTA Methods and Evaluation Criteria

We compare our EV-fusion with nine SOTA methods, including two AE-based methods, that is, Densefuse [13] and RFN-Nest [15]; two GAN-based methods, that is, FusionGAN [24] and GANMcC [27]; and five CNN-based methods, namely IFCNN [19], U2fusion [10], PIAFusion [32], SEAFusion [17], and DIVFusion [11].

To quantitatively evaluate different algorithms, we employ six evaluation metrics, which are visual information fidelity (VIF), average gradient (AG), entropy (EN), spatial frequency (SF), natural image quality evaluator (NIQE) [52], and color quality enhancement (CQE) [53]. VIF is a full-reference image quality assessment index based on natural scene statistics and human visual system, which has a good correlation with human judgment of visual quality. AG is used to measure the clarity of fused images. EN denotes the amount of information from the images. SF measures the spatial frequency. NIQE extracts features from natural landscapes to test the testing images, which are fit into a multivariate Gaussian model. CQE is a non-reference color quality measurement, which is based on the linear combination of colorfulness, sharpness, and contrast. Among them, a fusion method with higher VIF, AG, EN, SF, and CQE represents better fusion performance, while requiring lower NIQE.

C. Results and Discussion

1) Visual Comparison: Figs. 7–11 show the visual comparison of our EV-fusion with other nine SOTA methods on two different datasets. From the comparison results, it can be seen that all methods have certain image fusion effects. However, the AE-based methods, Densefuse and RFN-Nest, cannot extract the characteristics of different images well, with low brightness in the fused images, which looks similar to a weighted average of input images, such as pedestrian targets and car targets in Figs. 7, 8, and 10. The GAN-based method, FusionGAN, pay more attention on infrared images, resulting in blurry fusion results, with smooth edges and little background texture. As shown in Figs. 7 and 10, the edge of the pedestrian targets is unclear and the images are foggy. GANMcM has a better fusion effect than FusionGAN but loses many potential features in visible images. The fused results of U2fusion are foggy with low contrast, which is similar to FusionGAN. IFCNN, PIAFusion, and SeAFusion have good fusion performance but do not consider feature balance between infrared and visible images and ignore many potential features in visible images. As shown in Fig. 11, these

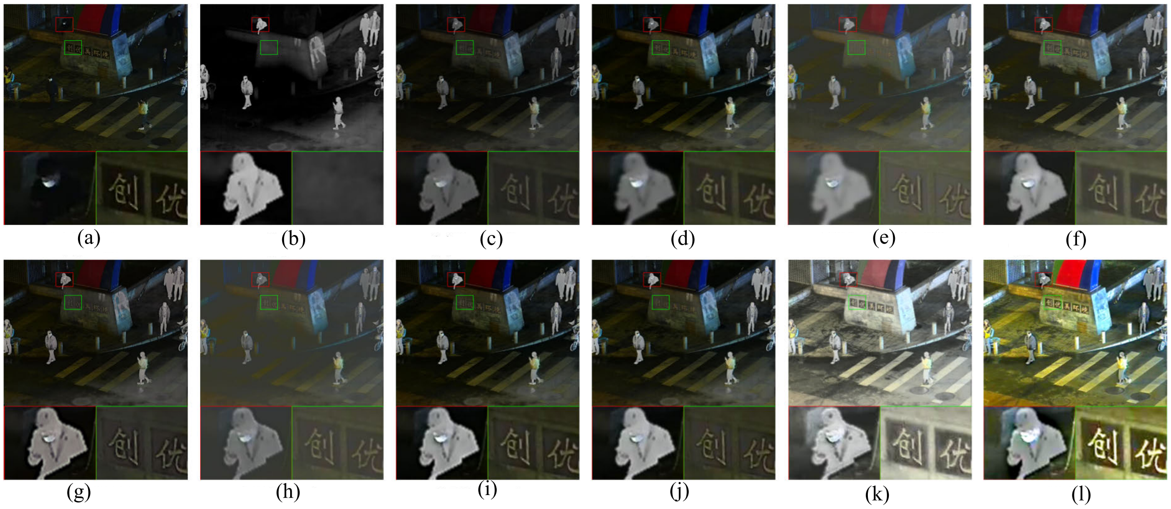


Fig. 7. Visual comparison of our EV-fusion with other nine SOTA methods on No. 010010 image in the LLVIP dataset. (a) VIS. (b) IR. (c) Densefuse. (d) RFN-Nest. (e) FusionGAN. (f) GANMcC. (g) IFCNN. (h) U2fusion. (i) PIAFusion. (j) SeAFusion. (k) DIVFusion. (l) Ours.

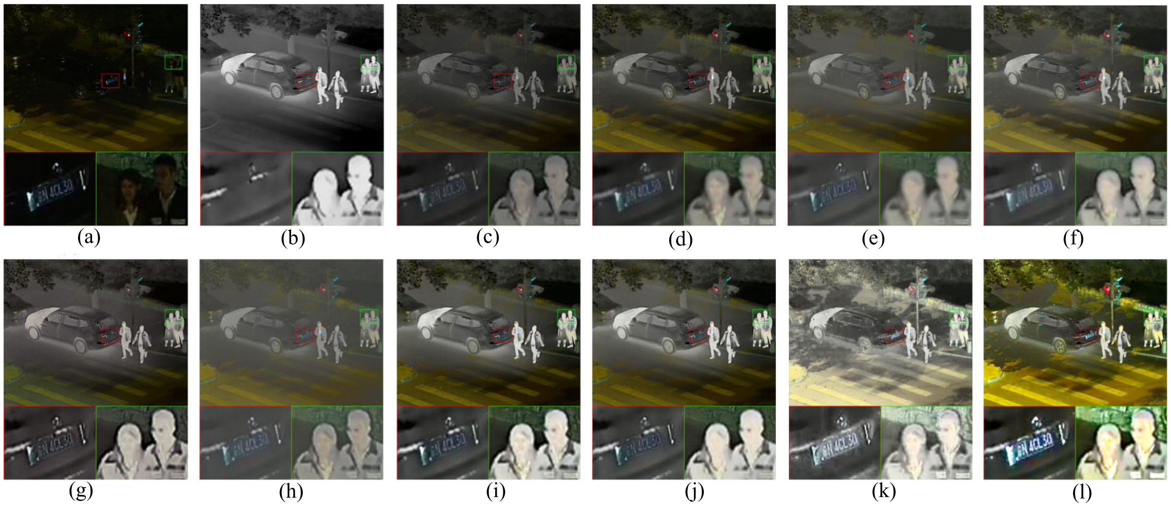


Fig. 8. Visual comparison of our EV-fusion with other nine SOTA methods on No. 090001 image in the LLVIP dataset. (a) VIS. (b) IR. (c) Densefuse. (d) RFN-Nest. (e) FusionGAN. (f) GANMcC. (g) IFCNN. (h) U2fusion. (i) PIAFusion. (j) SeAFusion. (k) DIVFusion. (l) Ours.

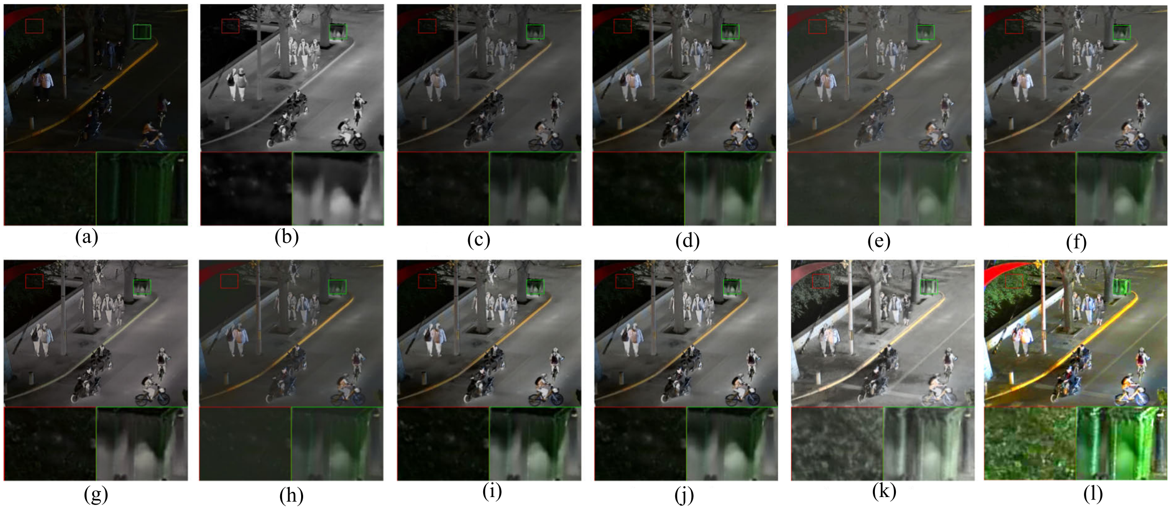


Fig. 9. Visual comparison of our EV-fusion with other nine SOTA methods on No. 040012 image in the LLVIP dataset. (a) VIS. (b) IR. (c) Densefuse. (d) RFN-Nest. (e) FusionGAN. (f) GANMcC. (g) IFCNN. (h) U2fusion. (i) PIAFusion. (j) SeAFusion. (k) DIVFusion. (l) Ours.

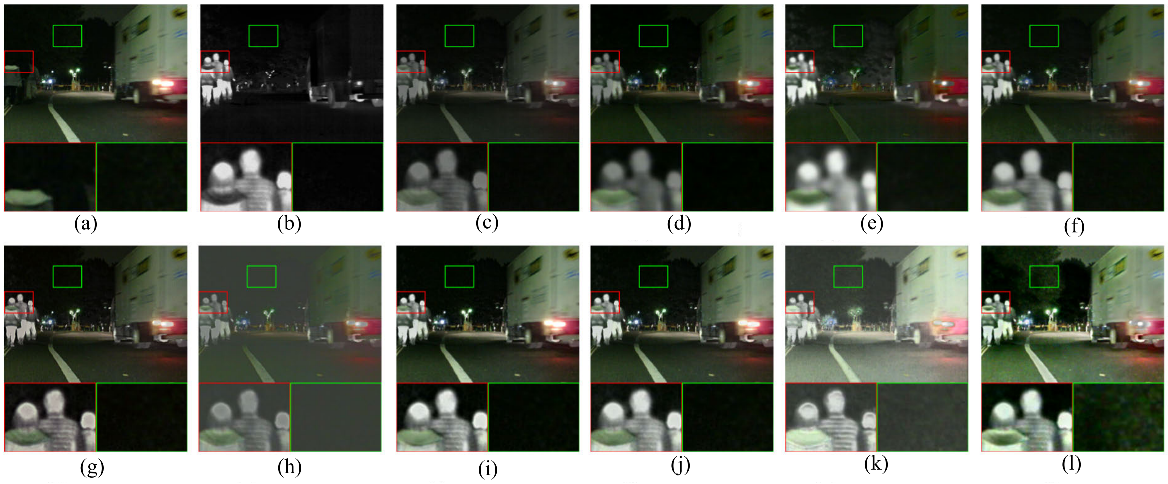


Fig. 10. Visual comparison of our EV-fusion with other nine SOTA methods on No. 00890N image in the MSRS dataset. (a) VIS. (b) IR. (c) Densefuse. (d) RFN-Nest. (e) FusionGAN. (f) GANMcC. (g) IFCNN. (h) U2fusion. (i) PIAFusion. (j) SeAFusion. (k) DIVFusion. (l) Ours.

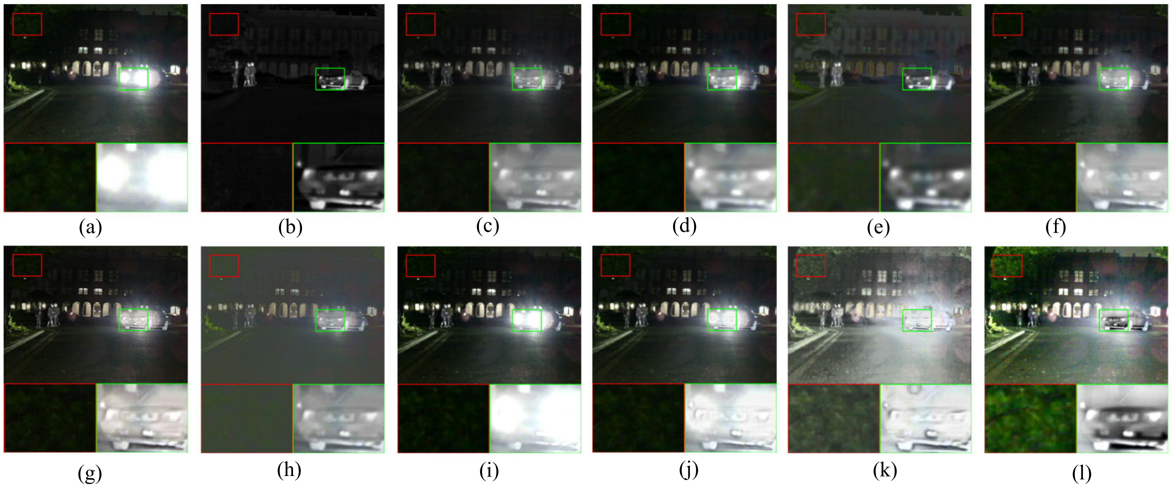


Fig. 11. Visual comparison of our EV-fusion with other nine SOTA methods on No. 00838N image in the MSRS dataset. (a)VIS. (b) IR. (c) Densefuse. (d) RFN-Nest. (e) FusionGAN. (f) GANMcC. (g) IFCNN. (h) U2fusion. (i) PIAFusion. (j) SeAFusion. (k) DIVFusion. (l) Ours.

methods retain too many features from visible image causing overexposure of car target areas, whereas our EV-fusion preserves detailed textures from infrared image. Among all compared methods, only DIVFusion introduces a visible image enhancement module, which has some brightness improvement effect. However, the fused results of DIVFusion are overall biased toward white, and the infrared target is not prominent. At the same time, noise interference and color distortion have also been introduced, which lead to poor visibility. In general, the results of our paper retain the prominent target features from infrared images while restoring potential information from visible images, which makes our fusion images have the best contrast, color fidelity, and visual effect.

2) Quantitative Comparison: We also conduct quantitative comparisons on two different datasets to verify the performance of our method. Tables I and II show the quantitative results of six objective metrics. It can be seen that our method ranks first in all indicators except for EN we rank second on both datasets. The best score of VIF and NIQE metric indicates that our method has the best visibility among all methods. The best score of AG and SF indicates that our results have more texture details. The best score of CQE demonstrates that our results have the best color fidelity. DIVFusion greatly improves the brightness of the results, which leads to a higher EN score. However, the contrast and color fidelity of its results are very poor. Figs. 12 and 13 show the cumulative distribution functions of six different metrics on two datasets, from which we can see the trend of these indicators across the entire dataset. Except for the EN indicator, our algorithm has significant advantages in all other indicators.

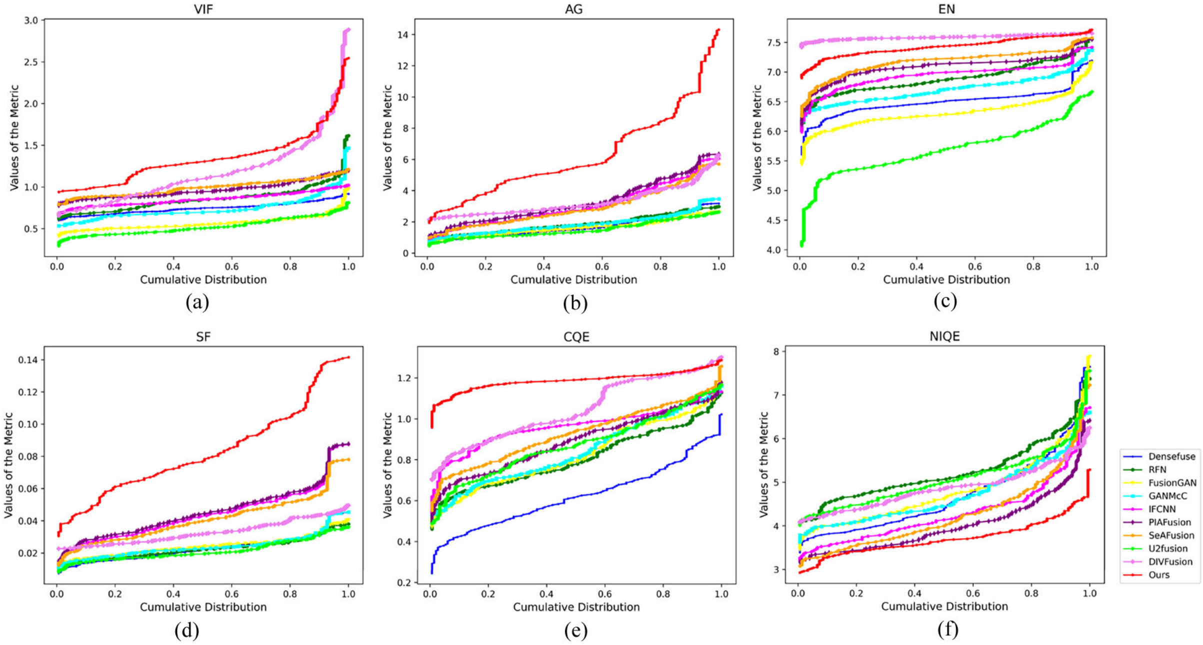
Overall, our method is able to extract potential features from visible images under dark environments and has better fusion performance compared to other SOTA methods in terms of both subjective and objective evaluation criteria.

D. Ablation Study

1) Comparison With Two-Stage Fusion Strategy: One common approach is to merge existing image enhancement models with image fusion models. However, we find that different models from different areas may have serious compatibility issue, directly coupling pre-trained enhancement algorithms

TABLE I
QUANTITATIVE COMPARISON WITH DIFFERENT METHODS OF LLVIP DATASET

Methods	VIF	AG	EN	SF	CQE	NIQE
Densefuse	0.7439	1.6556	6.4926	0.0228	0.6206	4.6982
RFN-Nest	0.8549	1.8447	6.8879	0.0227	0.8055	5.2317
FusionGAN	0.5630	1.6014	6.3075	0.0242	0.8234	4.7980
GANMcC	0.7371	1.8450	6.6995	0.0240	0.8390	4.7089
IFCNN	0.8519	3.0375	6.9365	0.0441	0.9587	4.2900
U2fusion	0.9618	3.2768	7.0831	0.0456	0.8863	4.0210
PIAFusion	0.9906	2.9052	7.1711	0.0409	0.9263	4.2494
SeAFusion	0.5232	1.4838	5.6887	0.0212	0.8701	5.0682
DIVFusion	1.1735	3.2522	**7.5857**	0.0326	1.0432	4.8598
ours	**1.3432**	**6.2156**	7.4239	**0.0828**	**1.1862**	**3.7063**



(a) (b) (c)

(d) (e) (f)

Fig. 12. Cumulative distribution of 6 metrics on 150 images from the LLVIP dataset. (a) VIF. (b) AG. (c) EN. (d) SF. (e) CQE. (f) NIQE.

TABLE II
QUANTITATIVE COMPARISON WITH DIFFERENT METHODS OF MSRS DATASET

Methods	VIF	AG	EN	SF	CQE	NIQE
Densefuse	0.7235	1.4821	5.4049	0.0182	0.6636	5.1295
RFN-Nest	0.7733	1.5821	5.7928	0.0212	0.7250	6.4626
FusionGAN	0.5304	1.3337	5.4628	0.0166	0.6711	5.7864
GANMcC	0.7346	1.9997	6.1304	0.0228	0.7560	4.9051
IFCNN	0.8814	2.6594	5.8373	0.0351	0.7781	3.9820
U2fusion	1.0274	2.8308	5.9598	0.0362	0.8031	4.0631
PIAFusion	0.9606	2.5771	6.0490	0.0328	0.7855	4.7569
SeAFusion	0.3204	1.2292	3.9696	0.0192	0.7168	5.7976
DIVFusion	0.8336	4.5061	**7.4985**	0.0454	1.0006	4.0358
ours	**1.1350**	**4.9683**	7.4478	**0.0577**	**1.0536**	**3.9092**

with fusion algorithms leads to poor visual effects in the fused images. To demonstrate the significance of our joint training model, we compared it with two-stage fusion strategy, which means the SOTA methods combined with low-light image enhancement methods. Low-light image enhancement is currently a popular research direction that can effectively restore lost details, textures, and color information in low-light images. In this article, we adopt the high-performance Zero-DCE for comparative experiments. To achieve a fair comparison, we retrain the Zero-DCE using the LLVIP dataset. In the experiment, visible images are first enhanced by the Zero-DCE algorithm, then the Y channel component is fused with infrared images by different fusion methods, and finally the enhanced color information is combined with the fused gray-scale images. As shown in Figs. 14 and 15, with the help of low-light image enhancement algorithm, other methods can better restore the features of the original image. However, the results of the Densefuse, IFCNN, and U2fusion suffer from

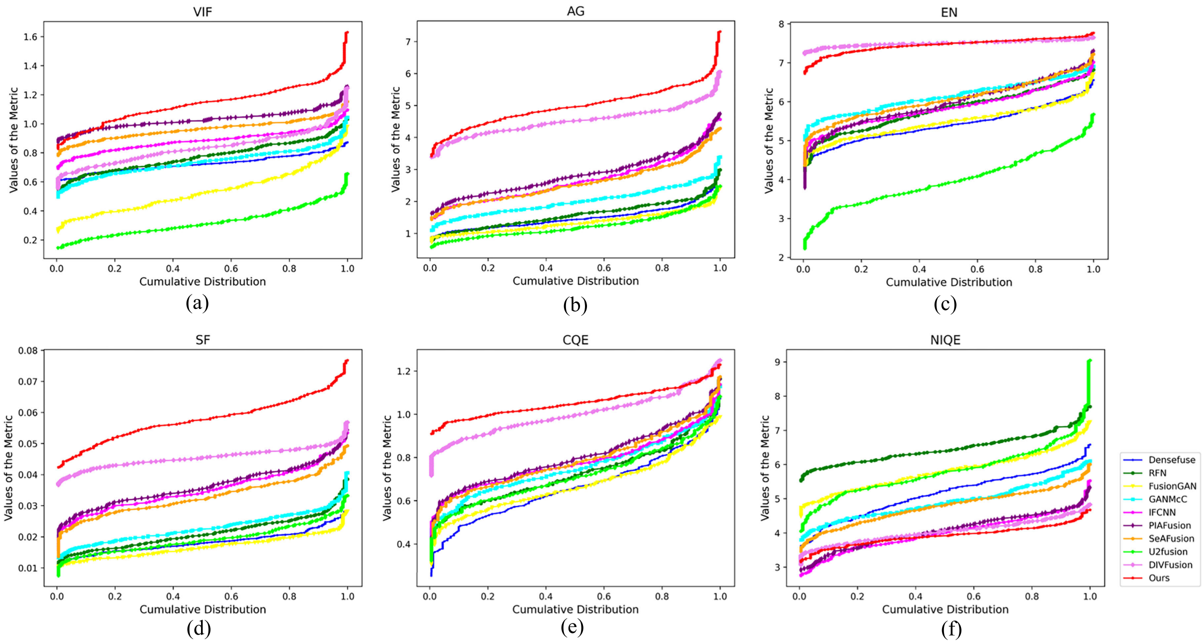


Fig. 13. Cumulative distribution of 6 metrics on 180 images from the MSRS dataset. (a) VIF. (b) AG. (c) EN. (d) SF. e CQE. (f) NIQE.

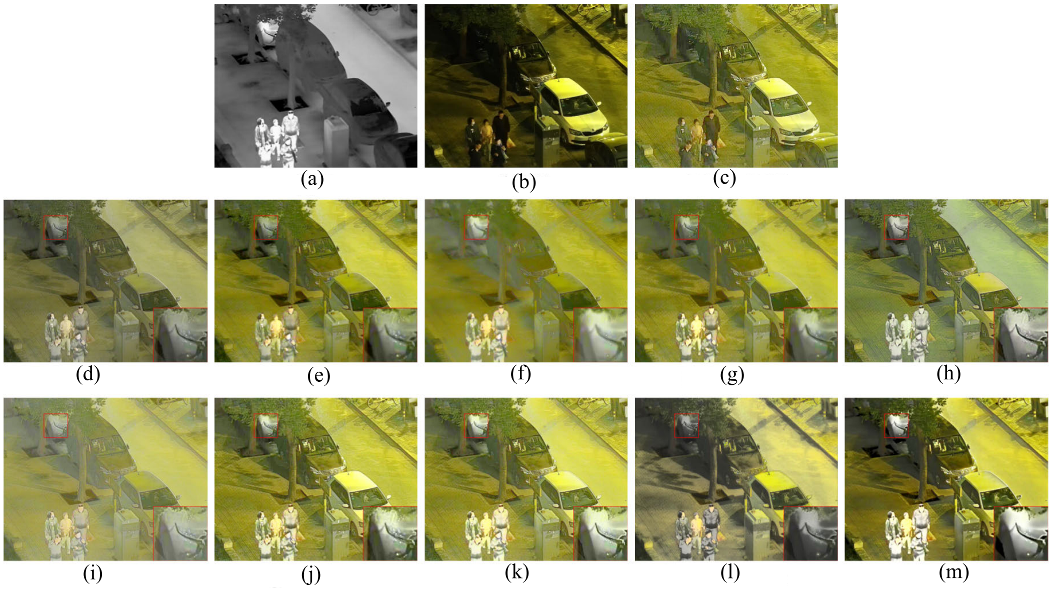


Fig. 14. Visual comparison of different methods integrated with Zero-DCE on No. 200008 image in the LLVIP dataset. (a) IR. (b) VIS. (c) Zero-DCE. (d) Densefuse. (e) RFN-Nest. (f) FusionGAN. (g) GANMcC. (h) IFCNN. (i) U2fusion. (j) PIAFusion. (k) SeAFusion. (l) DIVFusion. (m) Ours.

serious color distortion, which leads to poor visibility. The results of FusionGAN and GANMcC still contain too much information from infrared images and cause serious detail loss. The results of RFN-Nest, PIAFusion, and SeAFusion have great illumination, but low contrast. The results of DIVFusion have the worst color characteristics. In comparison, our method has the best brightness and contrast, making the overall image look more natural. Quantitative experimental results on LLVIP dataset are shown in Table III. Our method owns the best scores in VIF, AG, SF, and CQE metrics, while ranks second in EN and NIQE metrics. The results indicate that our method contains richer texture details and scene information compared with two-stage fusion strategy. Although the DIVfusion ranks first in EN and IFCNN ranks first in NIQE, DIVFusion causes serious color distortion and the results of the IFCNN have low contrast. Overall, compared to the two-stage fusion strategy, our method still has significant advantages with better visual effects for human perception.

2) Analysis of Bilateral-Guided Salience Map: The bilateral-guided salience map is introduced in fusion loss function by adopting bilateral-guided filtering for extracting infrared target regions and constraining weights of visible-light images in target areas. Fig. 16 shows the salience detection effect of

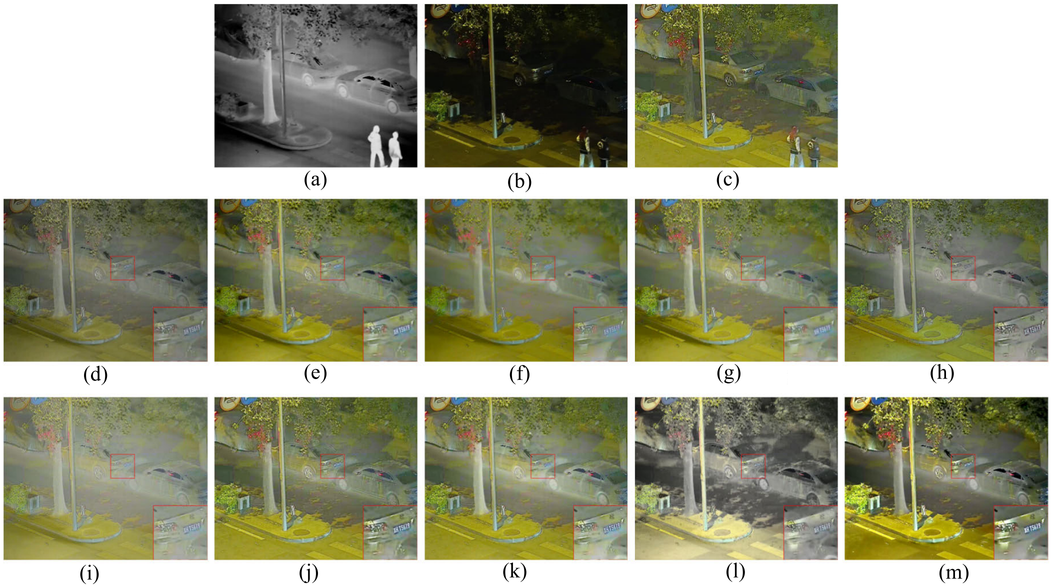


Fig. 15. Visual comparison of different methods integrated with Zero-DCE on No. 080004 image in the LLVIP dataset. (a) IR. (b) VIS. (c) Zero-DCE. (d) Densefuse. (e) RFN-Nest. (f) FusionGAN. (g) GANMcC. (h) IFCNN. (i) U2fusion. (j) PIAFusion. (k) SeAFusion. (l) DIVFusion. (m) Ours.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS INTEGRATED WITH ZERO-DCE ON LLVIP DATASET

Methods		VIF	AG	EN	SF	CQE	NIQE
Enhanced by Zero-DCE	Densefuse	0.7972	2.6857	6.7076	0.0319	0.7843	4.0383
	RFN-Nest	0.9088	3.0581	7.1789	0.0318	0.9352	4.4394
	FusionGAN	0.5659	2.6635	6.5470	0.0336	0.9129	4.1797
	GANMmC	0.8273	2.8936	6.9483	0.0335	0.9575	3.7850
	IFCNN	0.8625	5.1482	6.9240	0.0620	0.9323	**3.6917**
	U2fusion	1.0421	5.0948	6.9282	0.0617	0.9149	3.8830
	PIAFusion	0.9780	4.7485	7.0192	0.0568	0.9497	3.8050
	SeAFusion	0.6720	2.6408	6.3768	0.0295	0.9564	3.8294
DIVFusion		1.1735	3.2522	**7.5857**	0.0326	1.0432	4.8598
ours		**1.3432**	**6.2156**	7.4239	**0.0828**	**1.1862**	3.7063

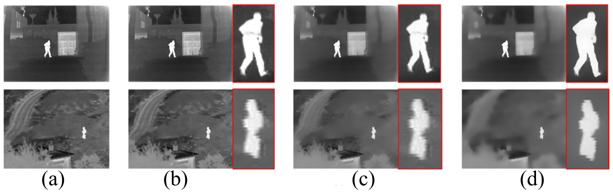


Fig. 16. Performance of the bilateral-guided filter. (a) Infrared image. (b) Bilateral filter. (c) Guided filter. (d) Bilateral-guided filter.

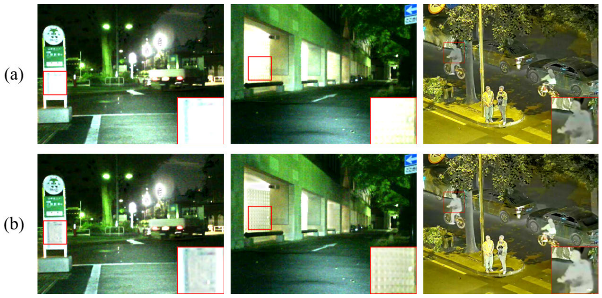


Fig. 17. Visualization of ablation study for bilateral-guided salience map. (a) w/o salience map. (b) EV-fusion.

bilateral-guided filtering. It can be seen that this filter can better extract the edges of the target and suppress background texture information compared to two separate filters. Therefore, it is necessary to verify the impact of this salience map on fusion performance. The results are shown in Fig. 17, with the help of salience map, the fused images not only show a significant improvement in brightness and contrast in the infrared target regions, but also effectively address the issue of overexposure. Quantitative results for salience map are shown in Table IV. For LLVIP dataset, the results without salience map achieve better values in AG and CQE metrics, which is reasonable. Due to the fact that fused images contain more infrared features in the infrared target areas, the gradient information of the image will be reduced to a certain extent. Under the influence of infrared intensity characteristics, the color of the targets becomes more white, resulting in some loss of color information. For MSRS dataset, the salience map helps our method achieve the best results in all metrics, except for CQE.

3) Analysis of Different Loss Functions: For the process of our joint training with enhancement and fusion tasks, the loss function is an important factor that affects the experimental results. In order to verify the necessity of each loss function,

TABLE IV
QUANTITATIVE RESULTS OF ABLATION STUDY FOR BILATERAL-GUIDED SALIENCE MAP. "W/O" MEANS WITHOUT

		VIF	AG	EN	SF	CQE	NIQE
LLVIP	w/o salience map	1.3218	**6.2267**	7.4109	0.0814	**1.1874**	3.7236
	ours	**1.3432**	6.2156	**7.4239**	**0.08289**	1.1862	**3.7063**
MSRS	w/o salience map	1.1271	4.7889	7.4170	0.0566	**1.0747**	3.9663
	ours	**1.1350**	**4.9683**	**7.4478**	**0.0577**	1.0536	**3.9092**

TABLE V
QUANTITATIVE RESULTS OF ABLATION STUDY FOR DIFFERENT LOSS FUNCTIONS ON LLVIP DATASET. "W/O" MEANS WITHOUT. BOLD INDICATES THE BEST AND UNDERLINED INDICATES THE SECOND BEST

		VIF	AG	EN	SF	CQE	NIQE
LLVIP	w/o illumination loss	0.9231	2.3559	6.6484	0.0407	1.1806	5.5319
	w/o color loss	1.3233	6.2039	**7.5188**	0.0803	1.1198	3.7324
	w/o smoothness loss	0.2321	**36.0476**	1.0036	0.4991	1.0199	20.7144
	w/omstructure loss	0.0440	0.8737	1.3788	0.0331	0.9731	18.8470
	w/o gradient loss	1.3070	5.7741	7.4008	0.0778	1.1540	3.8850
	ours	**1.3432**	6.2156	7.4239	**0.0828**	**1.1862**	**3.7063**
MSRS	w/o illumination loss	0.4529	1.6484	3.0003	0.0342	0.8029	7.4786
	w/o color loss	1.0965	4.7920	7.4313	0.0562	0.9997	3.9657
	w/o smoothness loss	0.1793	**54.1676**	1.0803	0.6153	0.8811	34.7084
	w/omstructure loss	0.0472	0.4837	1.2122	0.0173	0.5981	21.2640
	w/o gradient loss	1.1163	4.6717	7.4134	0.0555	1.0274	3.9094
	ours	**1.1350**	4.9683	**7.4478**	**0.0577**	**1.0536**	**3.9092**

TABLE VI
RUNNING TIME COMPARISON FOR DIFFERENT METHODS

	Densefuse	RFN-Nest	FusionGAN	GANMmC	IFCNN	PIAFusion	SeAFusion	U2fusion	DIVfusion	Ours
Time(s)	**0.011**	0.116	0.388	0.736	0.113	0.343	0.143	0.248	0.321	0.146

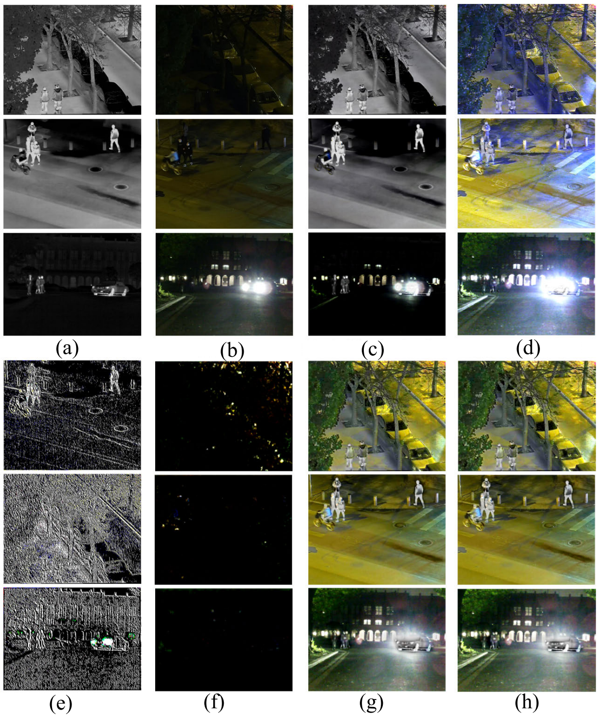


Fig. 18. Visualization of ablation study for different loss functions. (a) IR. (b) VIS. (c) w/o illumination loss. (d) w/o color loss. (e) w/o smoothness loss. (f) w/o structure loss. (g) w/o gradient loss. (h) Ours.

we conduct experiments on the proposed loss functions one by one. In the experiments, we retrain the network parameters after removing each loss function, respectively. Fig. 18 demonstrates the effectiveness of different loss functions. It is obvious that illumination loss, color loss, smoothness loss, and structure loss are essential. Without the constraint of illumination loss, the fusion result tends to be more biased toward infrared images. Without the constraint of color loss, the fused images suffer from serious color distortion. Without the constraint of smoothness loss, the results are all high-frequency information. Without the constraint of structure loss, the results are pitch black. The effect of gradient loss is not highlighted enough in the figure, but it can be seen from the quantitative indicators in Table V that the overall effect of fusion will be improved with the addition of gradient loss. Due to the high-frequency information presented in the results without smoothness loss, their AG is high, which is meaningless. As shown in Table V, under the assistance of all loss functions, our fusion effect has obvious advantages, quantitatively demonstrating the necessity of each loss function.

4) Computational Efficiency: Table VI denotes the computational efficiency of different methods for images with the size of 480 × 640. All the models are conducted on a server with Intel XEON GOLD 6226R and NVIDIA GeForce RTX3090. Our EV-fusion ranks moderate among all methods because of the introduction of visible image enhancement module and multihead self-attention block. Compared with DIVfusion, which also introduces an image enhancement module, our method shows great superiority.

V. Conclusion

The article proposes a novel image fusion model, called EV-fusion, for infrared and low-light color visible image fusion. To explore the detail and color features in the visible images captured under nighttime environment, we introduce an unsupervised visible image enhancement module based on several non-reference loss functions. We implement joint training for the image enhancement module and the image fusion module, simultaneously learning the correlation between the two tasks by optimizing their respective loss functions, which improves fusion performance. A bilateral-guided salience map by bilateral-guided filter is proposed in the fusion loss functions to improve the effect of infrared target regions in the fused results. We also design a SLGB in the fusion module to extract both local and global features from the source images. Extensive experiments on two datasets indicate the superiority of our method. Ablation studies of core factors in our model have demonstrated the effectiveness of our algorithmic innovation. In the future, we will further optimize the parameters and performance of the model, striving to achieve real-time high-contrast image fusion under nighttime environment.

References

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[2] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015, doi: 10.1016/j.inffus.2014.09.004.

[3] Z. Liu, Y. Feng, H. Chen, and L. Jiao, "A fusion algorithm for infrared and visible based on guided filtering and phase congruency in NSST domain," *Opt. Lasers Eng.*, vol. 97, pp. 71–77, Oct. 2017.

[4] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010, doi: 10.1109/TIM.2009.2026612.

[5] S. Zhang et al., "A multi-modal image fusion framework based on guided filter and sparse representation," *Opt. Lasers Eng.*, vol. 137, Feb. 2021, Art. no. 106354.

[6] N. Cvejic, J. Lewis, D. Bull, and N. Canagarajah, "Region-based multimodal image fusion using ICA bases," in *Proc. Int. Conf. Image Process.*, Mar. 2007, pp. 1801–1804, doi: 10.1109/icip.2006.312638.

[7] 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, doi: 10.1016/j.inffus.2016.02.001.

[8] Y. Chen, L. Cheng, H. Wu, F. Mo, and Z. Chen, "Infrared and visible image fusion based on iterative differential thermal information filter," *Opt. Lasers Eng.*, vol. 148, Jan. 2022, Art. no. 106776.

[9] 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.

[10] 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.

[11] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion," *Inf. Fusion*, vol. 91, pp. 477–493, Mar. 2023.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Y. Fu and X.-J. Wu, "A dual-branch network for infrared and visible image fusion," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10675–10680.

[17] 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.

[18] Z. Wang, W. Shao, Y. Chen, J. Xu, and L. Zhang, "A cross-scale iterative attentional adversarial fusion network for infrared and visible images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3677–3688, Aug. 2023.

[19] 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.

[20] Y. Yang et al., "Infrared and visible image fusion based on infrared background suppression," *Opt. Lasers Eng.*, vol. 164, May 2023, Art. no. 107528.

[21] 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.

[22] D. Zhu, W. Zhan, Y. Jiang, X. Xu, and R. Guo, "IPLF: A novel image pair learning fusion network for infrared and visible image," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8808–8817, May 2022.

[23] J. Zhang et al., "Multimodal image fusion via self-supervised transformer," *IEEE Sensors J.*, vol. 23, no. 9, pp. 9796–9807, May 2023.

[24] 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.

[25] Y. Gao, S. Ma, and J. Liu, "DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 549–561, Feb. 2023.

[26] 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.

[27] 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.

[28] J. Ma et al., "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.

[29] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[30] Q. Li et al., "Coupled GAN with relativistic discriminators for infrared and visible images fusion," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7458–7467, Mar. 2021.

[31] Z. Zhou, M. Dong, X. Xie, and Z. Gao, "Fusion of infrared and visible images for night-vision context enhancement," *Appl. Opt.*, vol. 55, no. 23, pp. 6480–6490, 2016.

[32] 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*, vol. 83, pp. 79–92, Jul. 2022.

[33] Y. Liu, L. Dong, and W. Xu, "Infrared and visible image fusion via salient object extraction and low-light region enhancement," *Infr. Phys. Technol.*, vol. 124, Aug. 2022, Art. no. 104223.

[34] D. Zou and B. Yang, "Infrared and low-light visible image fusion based on hybrid multiscale decomposition and adaptive light adjustment," *Opt. Lasers Eng.*, vol. 160, Jan. 2023, Art. no. 107268.

[35] S. Hao, T. He, B. An, X. Ma, H. Wen, and F. Wang, "VDFEFuse: A novel fusion approach to infrared and visible images," *Infr. Phys. Technol.*, vol. 121, Mar. 2022, Art. no. 104048.

[36] C. Guo et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1777–1786.

[37] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1632–1640.

[38] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," Aug. 2018, *arXiv:1808.04560*.

[39] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5627–5636.

[40] X. Zhang and X. Wang, "MARN: Multi-scale attention Retinex network for low-light image enhancement," *IEEE Access*, vol. 9, pp. 50939–50948, 2021.

[41] 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.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[44] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[45] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[46] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[47] X. Zhang, X. Wang, and C. Yan, "LL-CSFormer: A novel image denoiser for intensified CMOS sensing images under a low light environment," *Remote Sens.*, vol. 15, no. 10, p. 2483, May 2023.

[48] N. Zheng, J. Huang, F. Zhao, X. Fu, and F. Wu, "Unsupervised underexposed image enhancement via self-illuminated and perceptual guidance," *IEEE Trans. Multimedia*, vol. 25, pp. 5469–5484, Aug. 2022.

[49] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, Jul. 1980.

[50] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[51] 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. Workshops (ICCVW)*, Oct. 2021, pp. 3489–3497.

[52] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.

[53] K. Panetta, C. Gao, and S. Agaian, "No reference color image contrast and quality measures," *IEEE Trans. Consum. Electron.*, vol. 59, no. 3, pp. 643–651, Aug. 2013.

Xin Zhang received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Optics and Photonics.

His research interests include image restoration, including image fusion and low-light image enhancement.

Xia Wang received the Ph.D. degree in automation from the China University of Mining and Technology, Xuzhou, China, in 1999.

She is currently an Associate Professor with the Beijing Institute of Technology, Beijing, China, where she is also the Vice Dean of the Institute of Photoelectric Imaging and Information Engineering. Her current research interests include optoelectronic detection, spectrum analysis, and imaging technology.

Changda Yan received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Optics and Photonics.

His research interests include event camera, image processing, and visual navigation.

Qiyang Sun received the M.S. degree in engineering from Australia National University, Canberra, ACT, Australia, in 2012. He is currently pursuing the Ph.D. degree with the Electronic Information, Beijing Institute of Technology, Beijing, China.

His research interests include image processing, remote sensing, and 3-D environmental perception.