LLE-Fuse: Lightweight Infrared and Visible Light Image Fusion Based on Low-Light Image Enhancement

Song Qian , Liwei Yang , Yan Xue , Jiya Tian , Guzailinuer Yiming * , Junfei Yang

Posted Date: 28 June 2024

doi: 10.20944/preprints202406.2023.v1

Keywords: Image Fusion; Infrared and Visible Images; Low-Light Image Enhancement; Structural Re-parameterization; Lightweight Network.

Article

LLE-Fuse: Lightweight Infrared and Visible Light Image Fusion Based on Low-Light Image Enhancement

Song Qian [1], Liwei Yang [1,2], Yan Xue [1], Jiya Tian [1], Guzailinuer Yiming [1,*] and Junfei Yang [1]

[1] Xinjiang Institute of Technology, Faculty of Information Engineering, Aksu, 843100, China; song_qian@xjit.edu.cn (S.Q.); 18916336783ylw@gmail.com (L.Y); xue_yan@xjit.edu.cn (X.Y.); tianjiya@sohu.com (J.T.); Junfei_yang@xjit.edu.cn (J.Y)

[2] Shihezi University, Faculty of Mechanical Engineering, Shihezi 832003, China

* Correspondence: guzailinuer@xjit.edu.cn

Abstract: Infrared and visible light image fusion technology integrates image feature information from two different modalities to generate an image that integrates complementary information from the source images. However, in low-light scenarios, the degradation of lighting in visible light images leads to existing fusion methods being unable to effectively extract texture detail information from the scene. At this time, the target prominence information provided by infrared images alone is far from sufficient. To address this challenge, this paper proposes a lightweight infrared and visible light image fusion method based on low-light enhancement, named LLE-Fuse. Firstly, the method improves the MobileOne Block by embedding the Sobel operator in the Edge-MobileOne Block to perform feature extraction and downsampling on the source images. Then, intermediate features of different scales are obtained and fused through a cross-modal attention fusion module, followed by image reconstruction using a decoder with upsampling. Next, the CLAHE algorithm is used for image enhancement of infrared and visible light images, and the enhanced loss guides the network model to learn the capability of low-light enhancement. Finally, upon completion of the network training, the method optimizes the Edge-MobileOne Block into the direct connection structure of MobileNetV1 through structural re-parameterization, effectively reducing the computational resource consumption of the model network. Through extensive experimental comparisons, this paper's method demonstrates excellent performance in both visual appeal and evaluation metrics, providing satisfactory fusion results even in low-light scenarios.

Keywords: image fusion; infrared and visible images; low-light image enhancement; structural re-parameterization; lightweight network

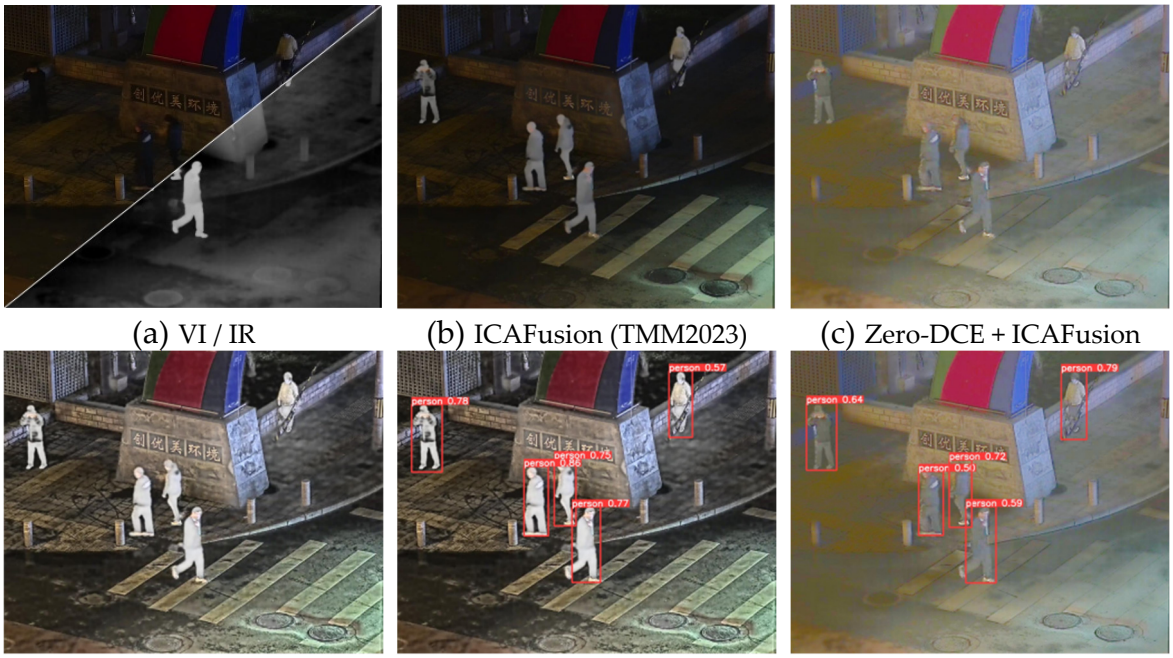
1. Introduction

Due to the differences in the working principles of sensors from various imaging modalities, the image scenes they capture exhibit unique characteristics [1]. Visible light images rely on capturing the light reflected by objects to form images, thus providing rich texture and structural information, as well as an intuitive visual experience. However, this characteristic of visible light images also makes them highly sensitive to environmental lighting and obstructions such as smoke, which can reduce the quality and usability of the images. In contrast, infrared images are formed by detecting the thermal radiation emitted by objects. This imaging method excels in highlighting targets and maintains stable image quality across various environments since it is not limited by external lighting conditions. Nevertheless, infrared images typically have lower resolution and lack detailed background texture information, which limits their performance in certain application scenarios. By fusing infrared and visible light images, the complementary advantages of both can be fully utilized to create an image that integrates information from both sources [2]. This fused image not only

 doi:10.20944/preprints202406.2023.v1

contains rich scene information but also excels in target recognition and detail presentation. In this way, the fused image can provide more comprehensive and accurate visual information, greatly enhancing the application value and practicality of the images.

Figure 1 presents an example of a low-light scenario. Figure 1a shows the visible and infrared images of the scene under low illumination. The result in Figure 1b is from an advanced image fusion method. In the field of image processing, particularly the fusion of infrared and visible light images under low-light conditions, presents a challenging problem. Low-light environments can severely affect the quality and details of visible light images, leading to a decline in visual perception and an inability to provide effective target and scene detail information for the fusion task. Even though the results shown in Figure 1c were obtained by first enhancing the low-light images using a leading image enhancement algorithm, Zero-DCE [3], and then performing image fusion, it is still difficult to achieve visually pleasing fusion results. Therefore, the combination of low-light enhancement and image fusion tasks remains a significant challenge. Figure 1d displays the result of the algorithm proposed in this paper, which offers a better visual perception. As shown in Figure 1e,f, the target detection algorithm YOLOv5s [4] performs detection on Figure 1e and 1c, respectively, with the result from this paper's algorithm better identifying pedestrians. An ideal image fusion method should not only pursue good visual effects but also ensure the integrity of the source image information and highlight significant targets, serving subsequent high-level visual tasks such as target detection. Only by meeting the needs of both human and machine visual perception can image fusion technology play a greater role in various practical applications, including pedestrian re-identification [5], target detection [6], and semantic segmentation [7].



(a) VI / IR (b) ICAFusion (TMM2023) (c) Zero-DCE + ICAFusion

(d) Ours (e) Ours + YOLOv5s (f) Zero-DCE + ICAFusion + YOLOv5s

Figure 1. Fusion and target detection results in the LLVIP's low-light scenario #010042 using ICAFusion and the method proposed in this paper.

In recent years, with the advancement of deep learning, image fusion methods based on deep learning architectures have gradually replaced traditional image fusion techniques such as multi-scale transformation [8], subspace transformation [9], sparse representation [10], and saliency analysis [11], achieving better visual performance in fusion results. These deep learning-based fusion methods are generally categorized into four types: autoencoder (AE)-based fusion methods, convolutional neural network (CNN)-based fusion methods, generative adversarial network (GAN)-based fusion methods, and unified architecture-based fusion methods. During the initial stages of

 doi:10.20944/preprints202406.2023.v1

integrating image fusion with deep learning, AE-based fusion methods utilized pre-trained encoders and decoders to accomplish feature extraction and image reconstruction, with feature fusion being carried out through specific fusion strategies. CNN-based fusion methods implement feature extraction, feature fusion, and image reconstruction in an end-to-end manner through meticulously designed network structures and loss functions, capable of generating unique fusion effects. GAN-based fusion methods build upon CNN-based fusion methods by introducing a generative adversarial mechanism, where the generator and discriminator engage in adversarial training. This allows the fusion results to closely approximate the probability distribution of the source images in an unsupervised setting.

Although existing deep learning-based fusion methods can effectively integrate the complementary information of infrared and visible light images, there are still several issues that require further resolution. Firstly, current fusion methods are designed under the assumption that texture detail information is derived from visible light images and that saliency information comes from infrared images. This assumption holds true under normal lighting conditions. However, in low-light night-time scenarios, the lack of environmental light in visible light images leads to severe degradation in image quality, causing much of the texture detail in the visible light images to be obscured by darkness. As a result, existing fusion methods struggle to effectively extract texture details, leading to a decline in the quality of the fusion results. A direct solution to this problem is to enhance visible light images using advanced low-light enhancement algorithms and then fuse the enhanced visible light images with infrared images. However, such methods often lead to compatibility issues. As shown in Figure 1c, the low-light enhancement algorithm alters the color distribution of light sources, exacerbating the problem to some extent. Moreover, stacking network models designed for different tasks is not conducive to the application and promotion of image fusion in practical scenarios. There is still a need to consider the design of lightweight networks further.

In response to the limitations of existing image fusion algorithms, this paper introduces a lightweight infrared and visible light image fusion network for low-light enhancement, called LLE-Fuse. This method utilizes a dual-branch architecture to extract features from infrared and visible light images separately and performs downsampling at various stages. It then fuses these features of different scales using a cross-modal attention fusion module. Following this, the integrated information is used to reconstruct the image through a network with upsampling. Finally, the CLAHE algorithm is applied for image enhancement of both infrared and visible light images, and the network model is guided to learn the capabilities of enhancing and fusing images in low-light scenarios through a combination of enhancement loss and correlation loss. Additionally, to further improve the network's representational capacity while maintaining its lightweight nature, this paper designs an Edge-MobileOne Block based on structural re-parameterization to serve as both the encoder and decoder. This enhances the fusion performance without increasing additional computational costs during the inference phase. The main contributions of this paper are as follows:

- This paper proposes a lightweight fusion framework for infrared and visible light images based on low-light enhancement, which can rapidly improve the visual perception of low-light scenes and produce fusion images with high contrast and clear textures;
- A novel enhancement loss is introduced in this paper to guide the network model in learning the intensity distribution and edge gradient of the CLAHE enhancement results, enabling the model to generate fusion images with low-light enhancement effects in an end-to-end manner;
- The paper designs the Edge-MobileOne Block as the encoder and decoder of the network model, significantly enhancing the fusion performance without increasing the computational load during the inference phase;
- The paper also designs a cross-modal attention fusion module that effectively integrates information from different modality images.

The remainder of this paper is organized as follows. In Section 2, a brief introduction to the related work on image fusion and semantic injection is provided. Section 3 presents a detailed description of the proposed LLE-Fuse, including the overall framework, Edge-MobileOne Block, cross-modal attention fusion module, and loss functions. Section 4 demonstrates, through extensive

 doi:10.20944/preprints202406.2023.v1

experimental comparisons, that our method outperforms other methods, which is followed by some concluding remarks in Section 5.

2. Related Work

2.1. Deep Learning-Based Fusion Methods

Autoencoder (AE)-based fusion methods utilize a pre-trained encoder and decoder for feature extraction and image reconstruction. They also employ manually designed fusion rules tailored to the characteristics of the source images to ensure the completion of the fusion task. DenseFuse [12] is a typical AE-based fusion method, which employs DenseNet [13] as its backbone network and uses addition and l1-Norm as fusion strategies. To further enhance the feature extraction capabilities of the autoencoder structure, Li et al. proposed NestFuse [14], designing a nested connection encoder and a spatial channel attention fusion module to encourage the network to obtain better detail features. To make the fusion strategy more suitable for specific fusion tasks, Li et al. also proposed RFN-Nest [15], which ensures rich texture details in the fusion results through residual fusion networks, detail preservation loss functions, and feature enhancement loss functions. Due to the lack of interpretability in fusion strategies, Xu et al. [16] introduced a learnable fusion rule to evaluate the importance of each pixel for classification results, thereby improving the interpretability of the fusion network. In addition, MUFusion proposed by Cheng et al. [17] mitigates the significant degradation problem in the training process through an adaptive loss that includes content loss and memory loss.

Convolutional neural network-based infrared and visible light image fusion methods achieve feature extraction, integration, and reconstruction of images by employing complex loss functions and network architectures. STDFusionNet [18] utilizes saliency target masks to assist in the fusion process, thereby better preserving the key features of the images. RXDNFuse [19] combines the structural advantages of ResNet [20] and DenseNet [13] to more comprehensively extract multi-scale features and achieve effective image fusion. Li et al. [21] adopt a meta-learning approach, which allows a single CNN model to complete image fusion at different resolutions, greatly enhancing the model's versatility. Zhu et al. [48], based on a set of gamma-corrected underexposed images, constructed a pixel-level weight map by analyzing global and local exposure to guide the fusion process, thereby achieving image fusion and defogging in adverse weather conditions. Liu et al. [49] proposed a probability-based method to verify the effectiveness of objective metrics in image fusion. Tang et al. [22] introduced a new semantic loss to make the fused images more suitable for high-level visual tasks. Subsequently, Tang et al. [23] re-examined the combination of image fusion models with high-level visual task models to further improve fusion performance. Therefore, the performance of image fusion should not only consider the visual effects and evaluation metrics of the fusion results but also meet the machine vision requirements for subsequent high-level visual tasks.

Considering the capability of generative adversarial networks (GANs) to estimate probability distributions in unsupervised learning, they have shown great potential in tasks such as image fusion. FusionGAN [24] is the first model to apply GAN technology to image fusion, creating a generative adversarial framework that allows the fused image to better capture textural structures. However, this single adversarial strategy may lead to modality imbalance in the fusion results. To address this challenge, Ma et al. [25] proposed a dual discriminator conditional generative adversarial network (DDcGAN), which involves both infrared and visible light images in the adversarial process, achieving a more balanced fusion effect. Building on DDcGAN, AttentionFGAN [26] introduced a multi-scale attention mechanism to better preserve foreground target information in infrared images and background details in visible light images. Although the training of dual discriminator GAN models has its complexities, Ma et al. [27] further proposed a GAN model with multi-class constraints to more effectively balance the information from infrared and visible light images. Despite the progress these methods have made in enhancing visual quality, they may not fully consider the impact of fusion results on subsequent high-level visual tasks. To solve this issue, Liu et al. [28] proposed a model that combines fusion and detection with a dual-layer optimization strategy. To achieve the desired fusion results, this paper meticulously designs a series of network structures and

 doi:10.20944/preprints202406.2023.v1

loss functions, implementing image fusion with low-light enhancement effects through CNN-based methods.

2.2. Low-Light Scenarios-Based Fusion Methods

Due to the limitations of lighting conditions in imaging scenarios, visible light images are prone to severe degradation in low-light night-time conditions, making it difficult for existing fusion methods to effectively extract the rich texture details from visible light images, resulting in poor quality of the fused result images. Currently, there is limited research within the field of image fusion on low-light image fusion. Although PIAFusion [29] takes lighting factors into account, its model is overly simplified and struggles to adapt to lighting adjustments in complex environments. Liu et al. [30] added a visual enhancement module behind the fusion network to mitigate the impact of unfavorable factors such as low light on images, thereby further improving the level of image information expression. Tang et al. [31] proposed the first scene illumination fusion architecture, providing more effective information for the fused image through low-light enhancement and dual-modality fusion, and improving the performance of subsequent visual tasks. Chang et al. [32] introduced an image decomposition network to guide the network in correcting the illumination component of the fused image, resulting in an enhanced fusion image.

Additionally, the aforementioned fusion methods tailored for low-light scenarios primarily achieve image fusion by connecting two distinct network pipelines, each designed for different tasks. However, this type of solution often employs complex network models and loss functions to constrain the final fusion outcome, potentially leading to increased computational demands. This could limit the practical application of image fusion in real-world scenarios. To tackle this challenge, this paper explicitly adopts the principle of lightweight models in network design, employing simpler and more effective approaches to accomplish the task of low-light image fusion.

3. Methods

3.1. Problem Formulation

When visible light images experience illumination degradation in low-light scenarios, existing fusion algorithms often fail to effectively extract the background texture details from the visible light images. Relying solely on the information from infrared images to compensate for this loss is insufficient. Therefore, how to effectively mine the information from the source images and guide the fusion network to generate higher quality fusion results has become the key to the fusion of infrared and visible light images at night. A simple and effective method is histogram equalization. Accordingly, in response to the aforementioned problem, this text employs the Contrast Limited Adaptive Histogram Equalization (CLAHE) method with low-light enhancement effects [33] to process both infrared and visible light images. The network model is guided by the loss function to generate fusion images with similar effects, allowing the network model to additionally learn the capability of low-light enhancement without increasing the inference burden.

To further elucidate the role of CLAHE in processing infrared and visible light images, this section also presents visual results from the perspectives of image intensity and edges. Specifically, given a pair of registered low-light scene infrared and visible light images, when calculating the intensity loss and gradient loss, the maximum intensity and maximum gradient results generated can be formulated as follows:

$$I_{int} = Max\left(I_{ir}, I_{vi}\right), \tag{1}$$

$$I_{grad} = Max\left(\nabla I_{ir}, \nabla I_{vi}\right), \tag{2}$$

In the equation, ∇ represents the Sobel operator, which is used to extract edge gradient information from images.

 doi:10.20944/preprints202406.2023.v1

Most fusion algorithms are designed based on the assumption that visible images contain a wealth of background texture details and that they have prominent target information. This assumption holds true in most imaging environments with sufficient lighting. However, in low-light scenarios, visible images suffer from severe degradation and are unable to provide effective texture detail information. Consequently, models trained under such conditions are inevitably incapable of generating high-quality fusion images.

To further exploit the information hidden in the darkness of low-light scene source images, a common approach is to enhance the visible light images using low-light enhancement methods. Contrast Limited Adaptive Histogram Equalization (CLAHE) is one of the simplest and most effective low-light enhancement methods. It can effectively improve the contrast of images and enhance their visual quality. After enhancing the infrared and visible light images with CLAHE, the calculation of the maximum intensity and maximum gradient results is then performed. This process is formulated as follows:

$$I_{int}^{en} = Max\left(I_{ir}^{en}, I_{vi}^{en}\right), \tag{3}$$

$$I_{grad}^{en} = Max\left(\nabla I_{ir}^{en}, \nabla I_{vi}^{en}\right), \tag{4}$$

Figure 2 provides an example of a typical low-light scenario, illustrating the comparison of intensity and gradient before and after processing infrared and visible light images with CLAHE. ∇ represents the image's edge gradient, 'En' denotes the image enhanced using the CLAHE, and $Max(\cdot)$ signifies the selection of the maximum pixel value.

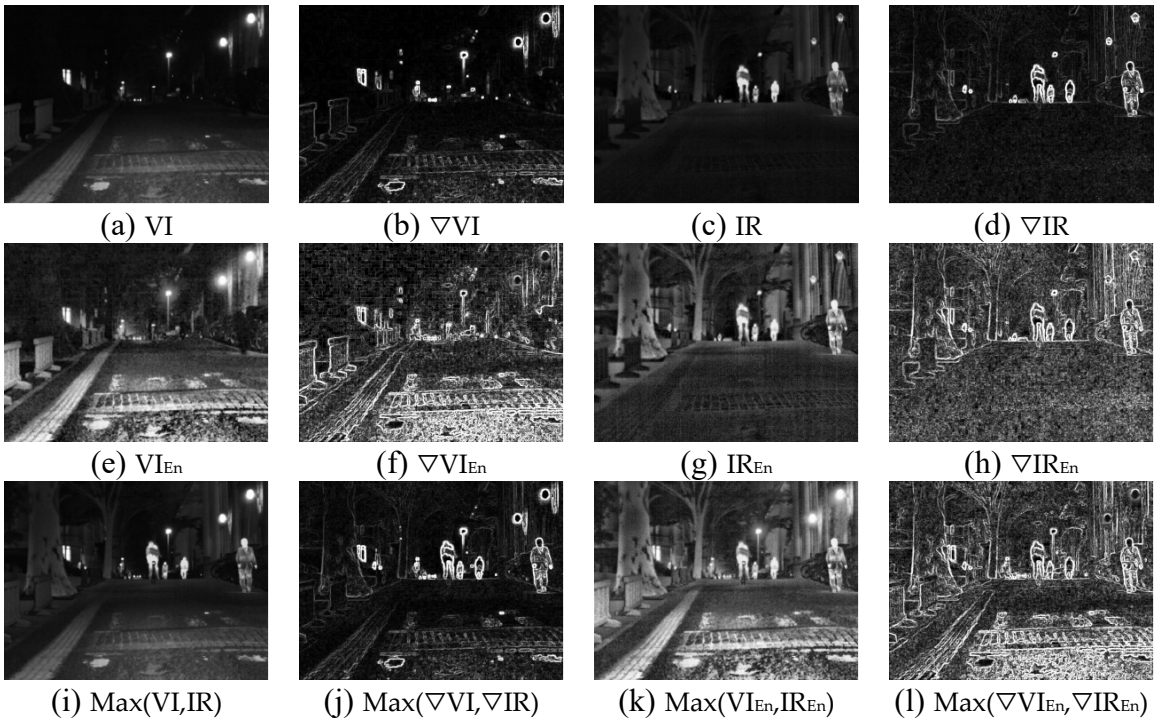


Figure 2. Intensity and Gradient Comparison of Infrared and Visible Light Images Before and After CLAHE Processing.

From Figure 2, it can be observed that both Figure 2e,g have better visual effects than Figure 2a,c, demonstrating that the CLAHE method achieves good results for both infrared and visible light images in low-light scenarios, and significantly enriches the texture detail information of the scene. Therefore, this paper will adopt the approach of Figure 2k,l as the pseudo-label images for the fusion network. Subsequently, by calculating the intensity distribution and edge gradient between the fusion image and the pseudo-label image through an enhancement loss, the fusion network can learn

 doi:10.20944/preprints202406.2023.v1

the low-light image enhancement capabilities similar to the CLAHE method. This allows the network to not only fuse the information of infrared and visible light images but also better address the image fusion issues in low-light scenarios. Moreover, this approach of constraining network training through a loss function enables the fusion network to possess low-light enhancement effects without additional computational overhead for the network model.

3.2. Network Architecture

3.2.1. Overall Network

To further enhance the representational capacity of the network model on the basis of lightweight requirements, the overall network framework of LLE-Fuse is shown in Figure 3, which can efficiently and concisely accomplish the task of image fusion.

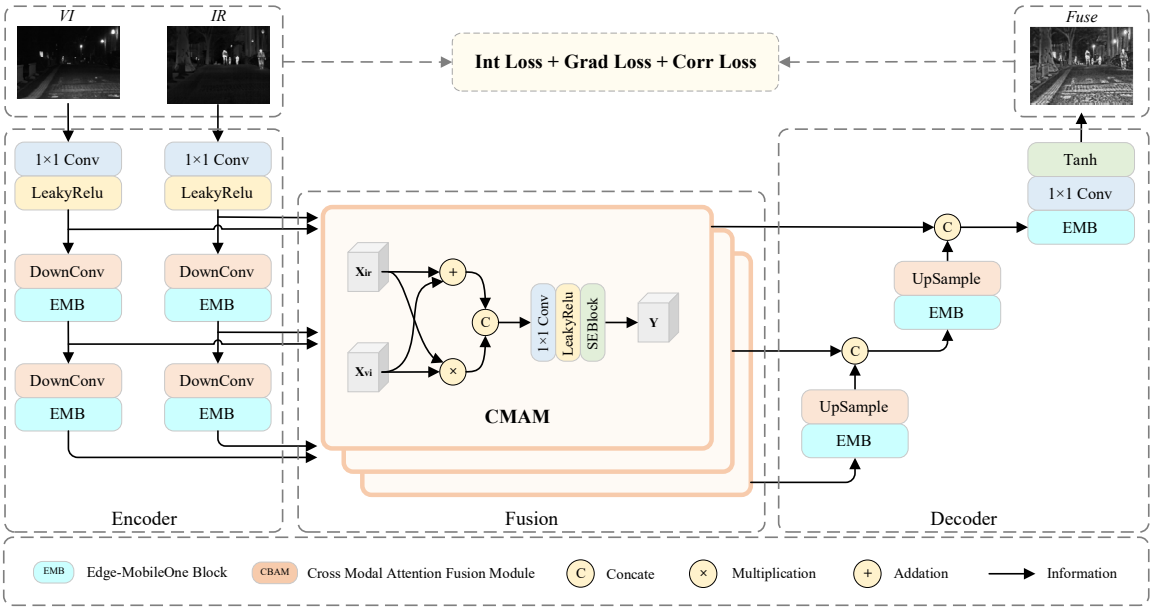


Figure 3. The overall network framework of LLE-Fuse comprises the Edge-MobileOne Block (EMB) and the Cross-Modal Attention Fusion Module (CMAM).

Specifically, during the training phase, infrared and visible light images are fed into different branches of the network to extract features. Initially, a 1x1 convolution is applied to extract shallow features, which are then passed into the EMB, where further extraction of texture and salient features is carried out. Throughout the feature extraction process, feature information from various stages is thoroughly integrated by the CMAM, combining both shared and unique features under a reduced computational load, effectively amalgamating the information from the source images. Following this, a decoder with an upsampling process integrates the features of different scales to reconstruct the image. Subsequently, CLAHE is utilized for low-light enhancement of both infrared and visible light images, aiding in the training of the fusion network through the loss function. Finally, upon completion of training, the entire network is optimized through structural re-parameterization, yielding the final network for the inference stage.

3.2.2. Edge-MobileOne Block

The model proposed in this paper is constructed from multiple Edge-MobileOne Blocks (EMB) that are improved based on the MobileOne Block [34], with the specific structure of this module shown in Figure 4. While standard convolutions have some effect in extracting features from infrared and visible light images, their feature extraction capabilities are relatively poor compared to complex models. However, replacing standard convolutions with complex modules often leads to an increased demand for computational resources in the network model. To address this issue, the

 doi:10.20944/preprints202406.2023.v1

method in this paper introduces structural re-parameterization technology to enrich the representational capacity of the network model without adding extra computational resource consumption during the inference phase.

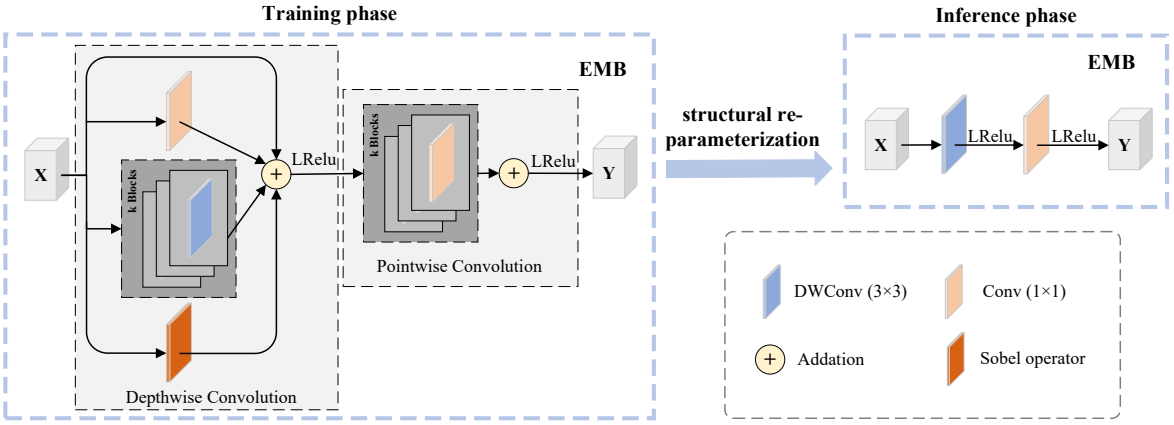


Figure 4. The network structure of the Edge-MobileOne Block (EMB).

Specifically, the EMB is primarily composed of a depthwise convolution section and a pointwise convolution section, with each section further expanded into k sub-branches, thereby implicitly enhancing the network's representational capacity. Through multiple experiments, k=4 has been selected as the parameter for the EMB. This process can be formulated as follows:

$$Y = Act\left(PConv_k\left(Act\left(DWConv_k\left(X\right)\right)\right)\right), k = \{1,2,3,4\}, \tag{5}$$

$$PConv\left(\cdot\right) = \sum_{k=1}^{n} Conv_{1\times1}^{k}\left(x\right), k = \{1,2,3,4\}, \tag{6}$$

$$DWConv\left(\cdot\right) = x + Conv_{1\times1}\left(x\right) + Sobel\left(x\right) + \sum_{k=1}^{n} Conv_{3\times3}^{k}\left(x\right), k = \{1,2,3,4\}, \tag{7}$$

In the formula, $Act\left(\cdot\right)$ represents the activation function LeakyReLU; $Sobel\left(\cdot\right)$ represents the Sobel edge gradient operator.

Strengthening image edge textures often effectively facilitates the progress of fusion tasks, but convolutional layers trained by the network also find it challenging to extract edge information parameters, necessitating the introduction of prior knowledge parameters capable of extracting edge information. Inspired by Chen et al. [35], an additional Sobel operator branch was incorporated into the depthwise convolution branch of the EMB to enhance the module's edge detection capabilities. It should be noted that the computation of this Sobel operator is consistent with that of DWConv. Therefore, the modified EMB, like the MobileOne Block, can be optimized into the network structure of MobileNetV1 through structural re-parameterization technology. Such an operation can further reduce the computational resource consumption of the model during the inference phase, and the performance improvement brought by the added Sobel branch is cost-free. After the equivalent transformation through structural re-parameterization technology, the processing procedure of the EMB is formulated as follows:

$$Y = Act\left(PConv'\left(Act\left(DWConv'\left(X\right)\right)\right)\right) \tag{8}$$

3.2.3. Cross-Modal Attention Fusion Module

 doi:10.20944/preprints202406.2023.v1

The fusion layer of our method is composed of the Cross-Modal Attention Fusion Module (CMAM), the specific structure of which is shown in Figure 3. Intermediate feature information at different scales from infrared and visible light images can be obtained through the encoder of the fusion network. Since these two features come from source images of different modalities, they have different focal points in the scene and contain complementary and common information between them. The fusion module should focus on the integration of complementary information from different modalities and the integration of common information. How to utilize the complementary information that exists only in one of the modalities to fuse these two features is a key issue. For thermal targets with sufficient lighting, both feature information should be enhanced during the fusion process. If methods similar to those used for processing complementary information are applied, it is likely that one of the feature information will be weakened.

To better address the issue of information fusion from different modalities, element-wise addition is used to extract the complementary information from the images of different modalities, while element-wise multiplication is employed to extract the common information from the images of different modalities. The element-wise addition and multiplication are represented as follows:

$$X_{add} = X_{vi} + X_{ir}, \tag{9}$$

$$X_{mul} = X_{vi} \times X_{ir}, \tag{10}$$

In the formula, X_{vi}, X_{ir} represents the depth features of infrared and visible light images extracted by the encoder. The element-wise addition X_{add} refers to handling the complementary information from different modalities through an element-wise addition operation on the visible light features and the thermal target features. On the other hand, the element-wise multiplication X_{mul} signifies strengthening the common information across different modalities by performing an element-wise multiplication operation on the visible light features and the thermal target features.

By utilizing CMAM (Channel and Modality Attention Module), the obtained common and complementary features are concatenated to form a feature vector that is then fed into a 1x1 convolutional layer to further compress the channel features. Subsequently, the Channel Attention Module (SEBlock) [36] enables the network to focus on the target regions, thereby enhancing the contrast of the targets. This process can be represented as:

$$Y = SE\left(Conv\left(Cat\left(X_{vi}, X_{ir}\right)\right)\right), \tag{11}$$

In the formula, $Cat(\cdot)$ denotes the feature cascading operation; $Conv(\cdot)$ represents the 1×1 convolutional block followed by the LeakyReLU activation function; $SE(\cdot)$ refers to the SEBlock, which is the channel attention module.

3.2.4. Color Space Model

To ensure that the fusion process retains the original color distribution as much as possible, the YCbCr color space model is introduced in this paper, as shown in Figure 5.

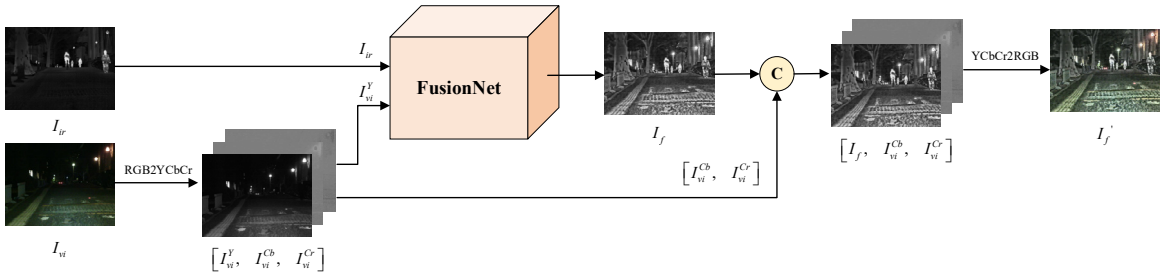


Figure 5. The color space model of LLE-Fuse.

 doi:10.20944/preprints202406.2023.v1

Specifically, the proposed algorithm converts the color space of the visible light image I_{vi} from RGB to YCbCr. Then, the luminance channel I_{vi}^{Y} and the infrared image I_{ir} are input into the fusion network to obtain the fusion result I_f. At this point, the fusion result is just a grayscale image with bright scenes but lacks color information. Therefore, the proposed algorithm replaces the luminance channel of the visible light image with the fusion result, and then converts the color space from YCbCr back to RGB, thus preserving the final color-combined image $I_f^{'}$. This method can avoid the problem of color distortion and better preserve the scene brightness of the fused image.

3.3. Loss Function

In the design of fusion networks based on convolutional neural networks, the selection and design of the loss function have a crucial impact on the network's performance. The loss function not only guides the weight updates during the network training process but also directly affects the feature representations learned by the network and the final output results. This paper employs an enhanced loss function L_{en} and a correlation loss function L_{corr} to jointly constrain the training of the network. The formula for the total loss is as follows:

$$Loss = L_{en} + \gamma L_{corr}, \tag{12}$$

In the formula, γ represents the weight coefficient for balancing the two losses.

The purpose of the enhanced loss is to guide the network model to learn the effect of low-light enhancement while maintaining high contrast and rich texture details. It is necessary for the model to learn based on both the intensity distribution and edge gradient of the fused images. The enhanced loss includes an intensity loss and a gradient loss, with the formulas as follows:

$$L_{en} = \alpha L_{Int} + \beta L_{Grad}, \tag{13}$$

In the formula, α and β are the weight coefficients for balancing these two losses.

To enable the model to better learn the effects of low-light enhancement, this paper employs CLAHE to improve the quality of both infrared and visible light images. This effectively enhances the local contrast of the images while also significantly reducing the noise level, resulting in processed images that exhibit superior visual effects, finer texture details, and more balanced contrast. Consequently, this paper will utilize CLAHE to enhance the visible light and infrared images, and collectively optimize the brightness and edge features of the fused images to achieve higher quality image fusion.

The intensity loss constrains the overall apparent intensity of the fused image. To allow the fused image to better possess the effects of low-light enhancement, this paper replaces the infrared and visible light images in the common intensity loss with the enhanced infrared and visible light images. The maximum intensity is then calculated to guide the training of the fusion network. The formula for the intensity loss is as follows:

$$L_{Int} = \frac{1}{HW}\left\| I_f - Max\left(I_{ir}^{en}, I_{vi}^{en}\right)\right\|_1, \tag{14}$$

In the formula, I_f represents the fused image; I_{ir}^{en}, I_{vi}^{en} represent the infrared and visible light images enhanced by the CLAHE algorithm, respectively; $Max(\cdot)$ denotes the selection of the maximum pixel value; H, W refers to the height and width of the input images; $\|\cdot\|_1$ signifies the *l1-Norm*.

To enrich the edge texture details of the fused image, the edge gradient of the fused image should be directed towards the maximum edge gradient of the enhanced infrared and visible light images. The formula for the edge gradient loss is as follows:

 doi:10.20944/preprints202406.2023.v1

$$L_{Grad} = \frac{1}{HW}\left\| \left|\nabla I_f\right| - Max\left(\left|\nabla I_{ir}^{en}\right|, \left|\nabla I_{vi}^{en}\right|\right)\right\|_1, \tag{15}$$

In the formula, ∇ represents the edge texture of the image, and in this chapter, the Sobel gradient operator is used to extract the edges of the image.

Furthermore, to better preserve the information of the source images, this paper also introduces a regularization term to strengthen the correlation between the fused image and the source images. The formula is as follows:

$$L_{Corr} = \frac{1}{corr\left(I_f, I_{ir}\right) + corr\left(I_f, I_{vi}\right)}, \tag{16}$$

In the formula, this paper uses $corr\left(x, y\right)$ to calculate the correlation between the fused image and the source image.

4. Experiments and Analysis

In this section, the paper first describes the experimental setup for network training, including the dataset, experimental details, comparative algorithms, and evaluation metrics. Then, comparative fusion experiments and generalization experiments are conducted to illustrate the superiority of the method proposed in this paper. Following that, an efficiency comparison experiment is carried out to demonstrate the lightweight nature of the method. Lastly, an ablation study is also performed to validate the effectiveness of the approach presented in this paper.

4.1. Experimental Setup

4.1.1. Dataset

To validate the effectiveness and generalizability of the proposed method in this paper, comparative experiments were conducted on three public datasets: LLVIP [37], TNO [38], and MSRS [29]. During the experimental training phase, the LLVIP dataset, which consists of registered infrared and visible light image pairs, was selected to train the model. For more convenient training, the image pairs in the training set were cropped to a size of 224 pixel × 224 pixel, and the CLAHE algorithm was applied to enhance the low-light conditions of both the cropped infrared and visible light images. In addition to conducting comparative experiments on the LLVIP dataset, this paper also carried out generalization experiments on the TNO and MSRS datasets to verify the performance of the proposed method.

4.1.2. Experimental Details

The method proposed in this paper is an end-to-end model, with the network optimizer being AdamW, epoch = 100, learning rate = 1×10^{-3}. The loss function parameters are denoted by $\alpha = 21, \beta = 43, \gamma = 5$. The testing set includes public datasets for the fusion of infrared and visible light images: LLVIP, TNO, and MSRS, with 50, 42, and 30 pairs of images selected for algorithm comparison experiments, respectively. The experiments in this paper were conducted on a GeForce RTX 2080Ti 11GB and an Intel Core i5-12600KF, using the deep learning framework PyTorch. All compared algorithms in the experiments were set according to their original papers.

4.1.3. Comparative Algorithms

To further validate the superiority of the algorithm proposed in this paper, LLE-Fuse is compared with 12 mainstream fusion methods, including three AE-based methods (DenseFuse [12], RFN-Nest [15], and CUFD [39]), six CNN-based methods (IFCNN [40], PMGI [41], SDNet [42], STDFusionNet [18], FLFuse [43], and U2Fusion [44]), and four GAN-based methods (FusionGAN [24], GANMcC [27], UMF-CMGR [45], and ICAFusion [46]). The performance of the algorithm in this

 doi:10.20944/preprints202406.2023.v1

paper is primarily measured through visual results and evaluation metric results in comparison with mainstream fusion methods.

4.1.4. Evaluation Metrics

Since the task of fusing infrared and visible light images does not have a reference image, a single evaluation metric is not sufficient to prove the superiority of the fusion effect. Therefore, this paper introduces five universal image quality evaluation metrics, namely Standard Deviation (SD), Visual Fidelity (VIF), Average Gradient (AG), Entropy (EN), and Spatial Frequency (SF), which measure the effectiveness of the fusion results from different perspectives. SD measures the high contrast of the fused image from the perspective of image contrast and distribution. VIF quantifies the amount of information shared between the fused image and the source image based on natural scene statistics and the human visual system, thereby measuring the degree to which the fusion result conforms to human visual perception. AG and SF measure the clarity of the fused image from gradient and frequency changes, respectively. EN assesses the amount of information contained in the fused image from the perspective of information quantity. These metrics are positive, meaning that the higher their values, the better the quality of the fused image. Through these comprehensive evaluation metrics, the performance of different fusion algorithms can be assessed and compared more objectively.

4.2. Comparative Experiments

The fusion performance for nighttime scenes is crucial for the fusion task. Therefore, this paper selects the LLVIP dataset, which is tailored for nighttime urban street scenes, for comparative experiments. Figures 6 and 7 display the visual results of the method proposed in this paper and 10 comparative algorithms on the LLVIP dataset. In these figures, background texture details are marked with green solid line frames, and infrared salient targets are marked with red solid line frames. To better showcase the details of the images, some of the solid line-framed images in this paper have also been enlarged.

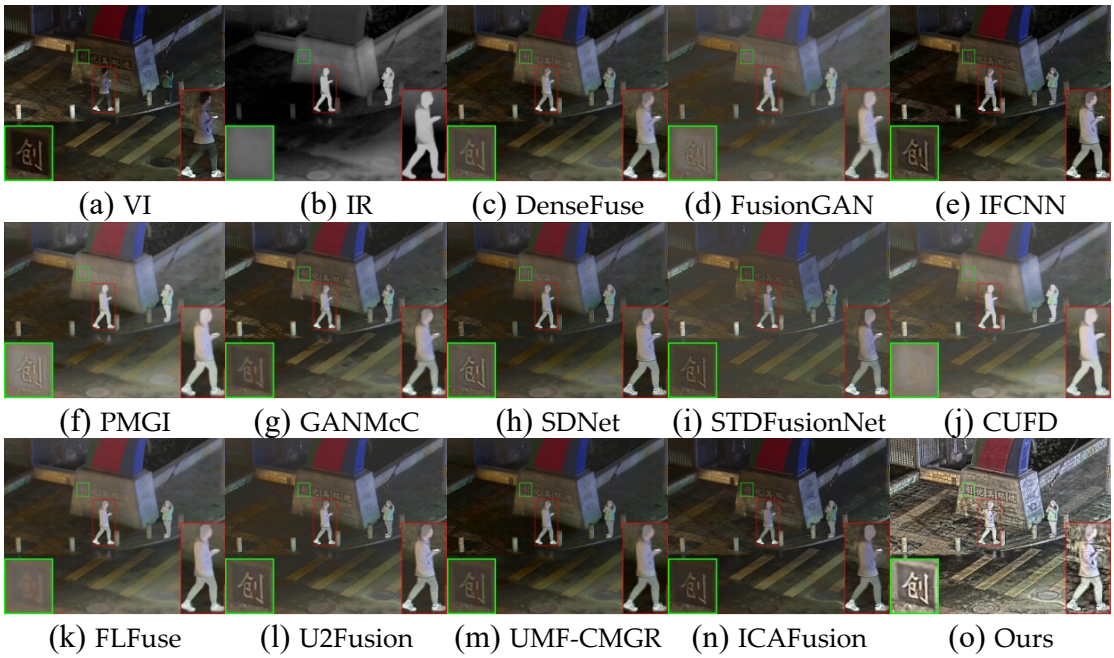


Figure 6. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the LLVIP dataset #010064.

 doi:10.20944/preprints202406.2023.v1

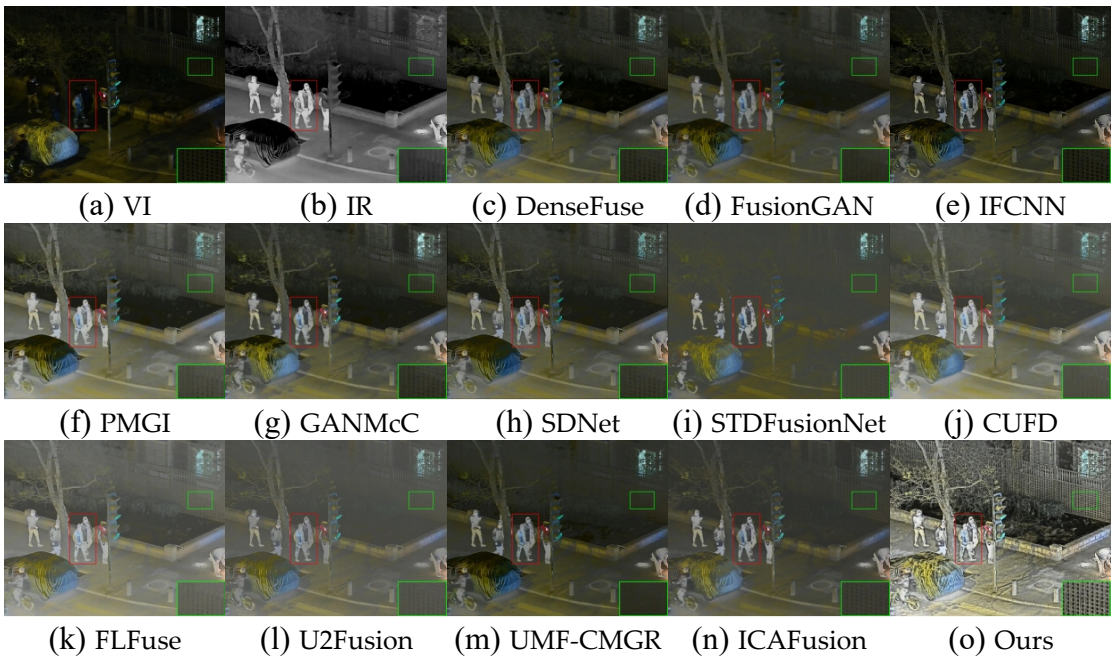


Figure 7. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the LLVIP dataset #060193.

In the comparative results shown in Figure 6, specific areas such as the text on the wall and the pedestrians in the scene have been zoomed in to examine the details of the fusion results in terms of background texture and target prominence. It can be observed that DenseFuse, GANMcC, SDNet, and UMF-CMGR exhibit average performance in terms of texture and prominence; FusionGAN loses some background texture details to a certain extent; IFCNN and ICAFusion do a good job of preserving the background texture of the visible light images but lack in prominence; PMGI and CUFD overall lean more towards the infrared image modality; STDFusionNet has insufficient contrast in the overall image; CUFD and FLFuse lose a significant amount of background texture details; the method proposed in this paper clearly demonstrates the effect of low-light enhancement, and the fusion results include more texture information from the visible light images while maintaining high contrast. Even in nighttime low-light scenarios, it can generate images akin to those captured during the daytime.

The comparative results in Figure 7 also readily demonstrate the superiority of the method proposed in this paper, further highlighting its advantages in terms of human visual perception. However, the comparison of visual results has a significant degree of subjectivity, and quantitative evaluation metrics are needed to further measure image quality from different aspects.

Table 1 presents the quantitative comparison results of the method proposed in this paper and 12 mainstream fusion methods within the LLVIP dataset. The data from Table 1 indicates that the method presented in this paper ranks first in all five metrics. The optimal SD suggests that the fusion results of our method encompass a richer set of contrast information. The values of AG and SF are significantly greater than those of other methods, indicating that the fusion results contain more detailed gradient texture information. This is attributed to the CLAHE in our method, which enhances the source images and strengthens the representation of texture details. The superior EN value compared to other methods indicates that the fusion results of our method can encapsulate a greater amount of information. A higher VIF score demonstrates that the fused image aligns more closely with the perceptual requirements of the human visual system, approximating the true image more closely. The significant advantages of our method across all evaluation metrics substantiate the superiority of our approach.

 doi:10.20944/preprints202406.2023.v1

Table 1. The quantitative comparison results in the LLVIP dataset. The best results are marked in bold, and the second-best results are underlined.

Dataset	Algorithm	SD	VIF	AG	EN	SF
LLVIP	DenseFuse	9.2490	0.8317	2.7245	6.8727	0.0363
	FusionGAN	8.3299	0.5322	1.9442	6.3078	0.0271
	IFCNN	8.6038	0.8094	4.1833	6.7336	0.0565
	PMGI	9.7091	0.7697	2.6571	6.9990	0.0332
	GANMcC	9.0244	0.7155	2.1196	6.6894	0.0267
	SDNet	8.9238	0.6537	3.4359	6.6793	0.0474
	STDFusionNet	6.2734	0.5222	2.9843	5.2143	0.0459
	CUFD	8.8130	0.7139	2.0277	6.6813	0.0284
	FLFuse	8.8942	0.6337	1.2916	6.4600	0.0162
	U2Fusion	7.7951	0.5631	2.2132	5.9464	0.0287
	UMF-CMGR	8.0539	0.5796	2.5040	6.4619	0.0389
	ICAFusion	7.8053	0.7300	2.4907	6.1222	0.0367
	Ours	**10.1525**	**1.1682**	**10.7486**	**7.6459**	**0.1122**

4.3. Generalization Comparative Experiments

The generalization ability of deep learning-based methods also reflects the performance of the model. Therefore, in addition to the comparative experiments on the LLVIP dataset, this paper also conducted experiments on the TNO and MSRS datasets. It is worth noting that the method proposed in this paper was trained on the LLVIP dataset and directly tested on the TNO and MSRS datasets.

4.3.1. TNO Dataset

To visually observe the fusion performance of different fusion algorithms on the TNO dataset, this paper selected two pairs of typical nighttime scene infrared and visible light images (#Kaptein_1123 and #Street), and the visual results are shown in Figures 8 and 9.

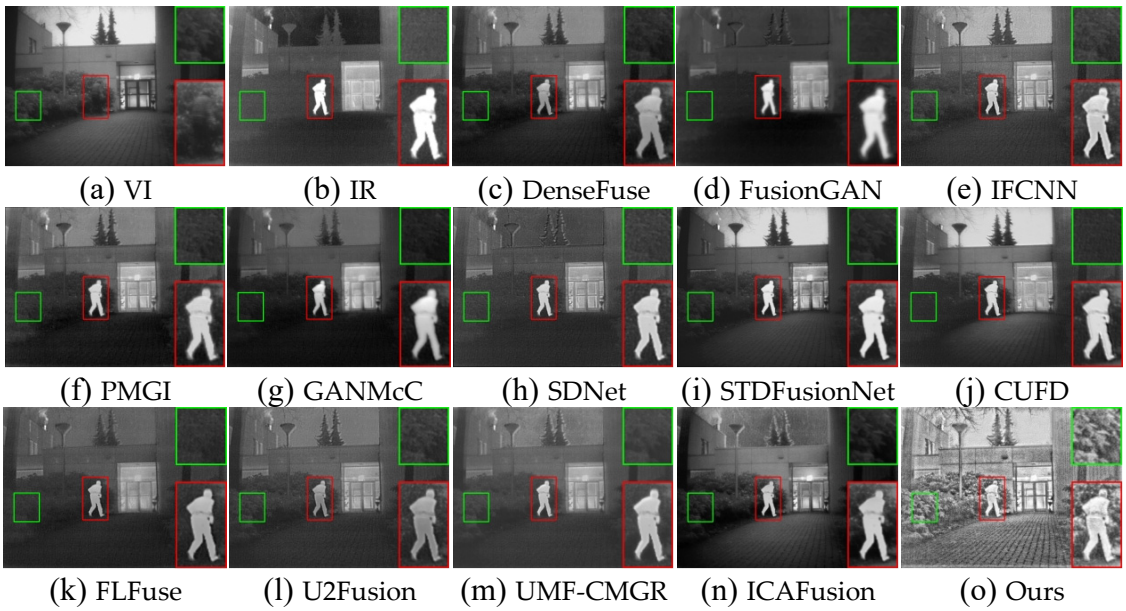


Figure 8. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the TNO dataset #Kaptein_1123.

 doi:10.20944/preprints202406.2023.v1

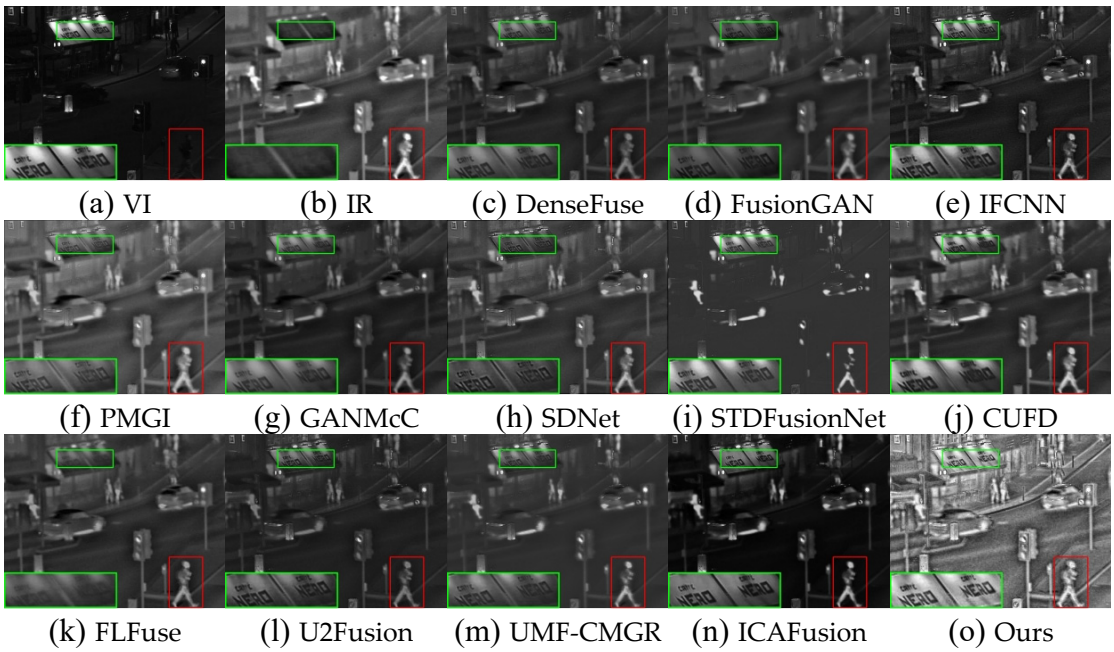


Figure 9. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the TNO dataset #Street.

From the visual results of Figures 8 and 9, it can be seen that the fusion results of FusionGAN, GANMcC, and UMF-CMGR are relatively blurry with a lack of scene details; DenseFuse, PMGI, SDNet, STDFusionNet, and U2Fusion have more background details, but the scenes are still limited by the illumination degradation of the visible light images, resulting in low contrast; while IFCNN, CUFD, and ICAFusion can provide better scene contrast, they still fail to offer a more visually appealing effect. In contrast, the method proposed in this paper still demonstrates surprising results on the TNO dataset, not only providing rich background texture details and salient thermal targets but also maintaining high contrast and illuminating the entire scene.

The quantitative comparison results of different methods on the TNO dataset are shown in Table 2. The data in Table 2 indicates that the method proposed in this paper performs the best overall on the TNO dataset, demonstrating a significant difference in multiple evaluation metric results. This is attributed to the low-light enhancement effect of the method proposed in this paper. Although our method lags behind others in the VIF metric, the visual images of the fusion results show that the fused images from our method are brighter and the scenes are clearer than those of other methods.

Table 2. The quantitative comparison results in the TNO dataset. The best results are marked in bold, and the second-best results are underlined.

Dataset	Algorithm	SD	VIF	AG	EN	SF
TNO	DenseFuse	9.2424	0.8175	3.5600	6.8193	0.0352
	FusionGAN	8.6736	0.6541	2.4211	6.5580	0.0246
	IFCNN	9.0581	0.7864	<u>5.1154</u>	6.8533	<u>0.0508</u>
	PMGI	<u>9.6029</u>	0.8689	3.6004	7.0181	0.0344
	GANMcC	9.0532	0.7123	2.5441	6.7359	0.0242
	SDNet	9.0698	0.7592	4.6117	6.6948	0.0457
	STDFusionNet	9.0451	0.9746	4.3846	6.9031	0.0455
	CUFD	9.5380	<u>0.9911</u>	4.0435	7.0599	0.0391
	FLFuse	9.2611	0.8084	3.3691	6.3924	0.0339
	U2Fusion	8.8553	0.6787	3.4891	6.4230	0.0327

 doi:10.20944/preprints202406.2023.v1

UMF-CMGR	8.7085	0.7121	2.9727	6.5325	0.0321
ICAFusion	9.5750	**1.0757**	4.6253	7.1372	0.0470
Ours	**10.3190**	0.7849	**13.3298**	**7.3418**	**0.1218**

4.3.2. MSRS Dataset

The MSRS dataset also includes some low-light scenes with insufficient illumination. This paper has selected two typical examples, and the visual results of different algorithms are shown in Figures 10 and 11. The method proposed in this paper is capable of providing brighter scenes with prominent targets and fully explores the background information hidden in the darkness.

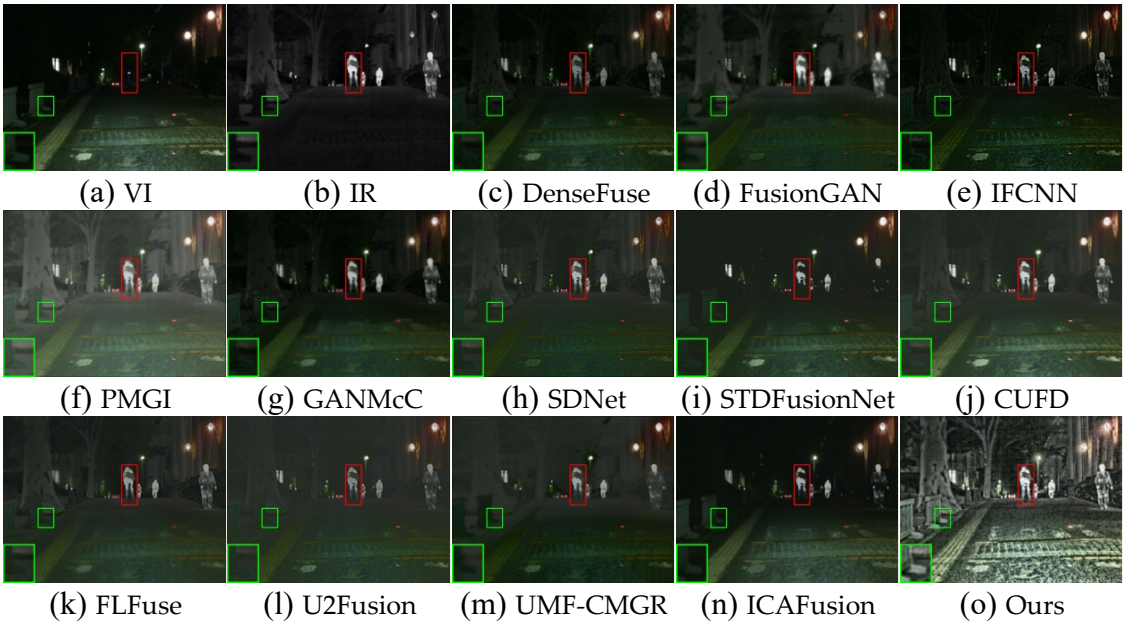


Figure 10. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the MSRS dataset #01023N.

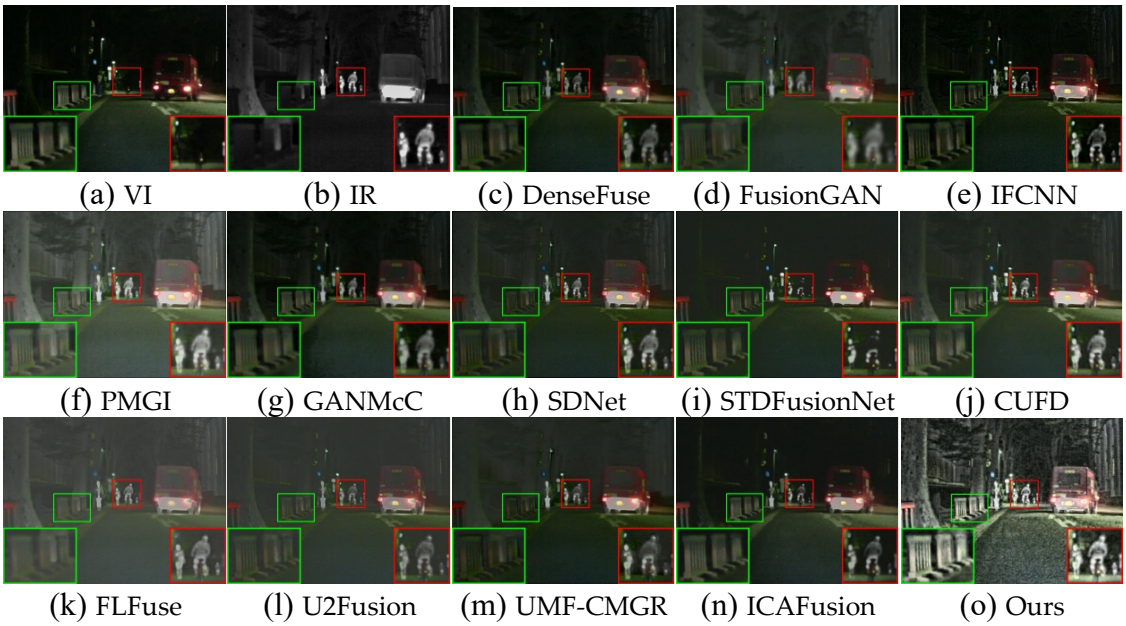


Figure 11. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the MSRS dataset #01037N.

 doi:10.20944/preprints202406.2023.v1

As shown in Figure 10 and 11, although existing mainstream fusion methods maintain the information of the source images to varying degrees, only the method in this paper enhances the image contrast of the entire scene. Compared to the source images, the fusion results of this paper have richer texture details and a better visual experience, as bright as daytime scenes.

The quantitative comparison results in the MSRS dataset are shown in Table 3. According to the results in Table 3, the LLE-Fuse method proposed in this paper exhibits a clear advantage across all five evaluation metrics. Specifically, the top ranking in the SD metric indicates that this method can produce fusion images with high contrast. The first place in the VIF metric suggests that the fusion results are more in line with the human visual perception system. The first place in the EN metric indicates that this method can contain rich information. The top rankings in the AG and SF metrics demonstrate that this method can fully preserve the texture information in the source images from both gradient and frequency perspectives. In summary, both qualitative and quantitative analysis results confirm the superiority of LLE-Fuse on the MSRS dataset.

Table 3. The quantitative comparison results in the MSRS dataset. The best results are marked in bold, and the second-best results are underlined.

Dataset	Algorithm	SD	VIF	AG	EN	SF
MSRS	DenseFuse	7.5090	<u>0.7317</u>	2.2024	6.0225	0.0255
	FusionGAN	5.7942	0.4671	1.4470	5.4631	0.0172
	IFCNN	6.6247	0.6904	<u>3.6574</u>	5.8457	<u>0.0450</u>
	PMGI	7.5838	0.6348	2.7487	<u>6.1807</u>	0.0301
	GANMcC	<u>8.0840</u>	0.6283	1.9036	6.0204	0.0212
	SDNet	5.6207	0.4149	2.5085	5.1713	0.0317
	STDFusionNet	6.5162	0.5298	2.4355	5.3721	0.0366
	CUFD	7.2588	0.6038	2.4734	5.8668	0.0322
	FLFuse	6.6117	0.4791	1.7241	5.5299	0.0189
	U2Fusion	5.7280	0.3902	1.8871	4.7535	0.0243
	UMF-CMGR	5.9766	0.3836	2.1143	5.5499	0.0272
	ICAFusion	7.8528	0.5961	1.9544	5.7254	0.0261
	Ours	**9.3456**	**0.9212**	**8.9116**	**7.3362**	**0.0789**

4.4. Efficiency Comparative Experiment

This paper conducted a runtime test on 50 images of the LLVIP dataset with a resolution of 1280×1024 to further evaluate the operational efficiency of the proposed algorithm compared to mainstream fusion algorithms. The average comparison results of the runtime are shown in the Table 4.

Table 4. The operational efficiency comparison results of the algorithm proposed in this paper and 12 mainstream algorithms on the LLVIP dataset.

Algorithm	Running Time(ms)	FLOPs/G
DenseFuse	267.26 ± 472.11	115.3
FusionGAN	341.21 ± 546.12	1213.9
IFCNN	20.42 ± 2.13	170.9
PMGI	218.26 ± 306.99	55.3
GANMcC	726.51 ± 972.39	2444.6
SDNet	111.22 ± 338.89	122.7
STDFusionNet	284.86 ± 517.61	369.5

 doi:10.20944/preprints202406.2023.v1

CUFD	No Data[1]	137.7
FLFuse	**1.77 ± 14.18**	18.6
U2Fusion	511.95 ± 774.04	863.4
UMF-CMGR	260.55 ± 440.31	824.3
ICAFusion	1456.16 ± 395.66	1226.2
Ours* [2]	6.43 ± 3.89	16.5
Ours	3.27 ± 3.54	**7.9**

[1] Since the average runtime of CUFD exceeds 5 seconds, it is not included in the statistics. [2] Ours* is LLE-Fuse without re-parameterisation.

As the table indicates, although the average runtime of the algorithm proposed in this paper is slightly worse than that of FLFuse, the smaller standard deviation suggests that the method is more stable. By comparing FLOPs, the algorithm presented in this paper exhibits better operational efficiency. Moreover, the fusion performance of the proposed method is significantly better than that of FLFuse, making this minor difference acceptable. The algorithm presented in this paper is based on the concept of structural re-parametrization and is equivalently optimized into a MobileNetV1 [47] structure during the inference phase, which can greatly reduce the fusion time.

Additionally, this paper compared the runtime, forward propagation memory requirements, number of parameters, weight size, and per-pixel cumulative deviation of the fusion results before and after structural re-parametrization on the LLVIP dataset. The comparison results of the model parameters are shown in Table 5. According to the data in Table 5, it can be seen that the method proposed in this paper can effectively reduce the computational resource consumption of the network through structural re-parametrization techniques while maintaining the same fusion effect.

Table 5. Comparison of model parameters before and after structural re-parametrization.

Re-parametrization	Runtime	Forward memory	Parameters	Weight Size	Deviation
×	6.43ms	14770.00MB	82,081	0.31MB	/
√	3.27ms	5620.18MB	47,089	0.18MB	2.8×10^{-3}

4.5. Ablation study

To further validate the effectiveness of the various modules designed in the method proposed in this paper, an ablation study was also conducted. The method proposed in this paper is named Experiment 1, the version without the enhancement loss is named Experiment 2, the method derived from Experiment 2 without the CMAM is named Experiment 3, and the method derived from Experiment 3 without the EMB is named Experiment 4. The qualitative and quantitative results of the ablation study are shown in Figure 12 and Table 6, respectively.

(a) VI/IR (b) Experiment1 (c) Experiment2 (d) Experiment3 (e) Experiment4

Figure 12. The ablation experimental results of the algorithm proposed in this paper on the LLVIP dataset #010177 and #260092.

 doi:10.20944/preprints202406.2023.v1

Table 6. The quantitative evaluation metric results of the ablation experiments on the LLVIP dataset. The best results are marked in bold, and the second-best results are underlined.

Experiment				Evaluation Methods				
No.	**EMB**	**CMAM**	L_{en}	**SD**	**VIF**	**AG**	**EN**	**SF**
1	√	√	√	**10.1213**	**1.1624**	**11.3551**	**7.6509**	**0.1183**
2	√	√	×	<u>9.8140</u>	0.9353	<u>7.6129</u>	<u>7.4839</u>	0.0771
3	√	×	×	9.5538	<u>1.0152</u>	4.0657	7.3336	0.0543
4	×	×	×	9.0162	0.8351	6.3087	7.0437	<u>0.0780</u>

4.5.1. Analysis of Enhancement Loss (L_{en})

The enhancement loss guides the network to generate fusion images with low-light enhancement effects. Without the enhancement loss, it becomes challenging for the fusion images to express complete scene information, and rich texture details are hidden in the darkness. In Figure 11, although Experiment 2 can produce a decent fusion image, compared to the results of Experiment 1, the scene in Experiment 2's result is noticeably darker, and the texture details are relatively fewer. Additionally, the differences in evaluation metrics also demonstrate the effectiveness of the enhancement loss, which can significantly improve the overall contrast of the fusion image and maintain richer textures.

4.5.2. Analysis of Cross-Modality Attention Fusion Module (CMAM)

The method proposed in this paper introduces a Cross-Modality Attention Fusion Module (CMAM) for fusing feature information from images of different modalities. As shown in Figure 11d, when CMAM is removed on the basis of Experiment 2, the prominence of the fusion result is reduced. Moreover, the decline in multiple evaluation metrics to varying degrees also indicates the effectiveness of this module.

4.5.3. Analysis of Edge-MobileOne Block (EMB)

The method proposed in this paper introduces an Edge-MobileOne Block (EMB) for extracting feature information from images of different modalities. As shown in Figure 11e, when EMB is removed on the basis of Experiment 3, the fusion network fails to effectively extract information from the source images, and the overall scene leans more towards the visible light image modality. Furthermore, the decrease in the values of the evaluation metrics in Table 6 further indicates a deterioration in performance. In summary, both qualitative and quantitative results show the role of this module in the overall network.

5. Conclusions

In this study, a novel lightweight method for fusing infrared and visible light images named LLE-Fuse, which is based on low-light enhancement, has been proposed. This method is an improvement over MobileOne and U-Net, utilizing an Edge-MobileOne Block embedded with the Sobel operator to extract information from both infrared and visible light images. It fuses these heterogeneous feature information through a Cross-Modality Attention Fusion Module and reconstructs the image using the decoder in the upsampling process. Additionally, the CLAHE method is applied as a preprocessing step for low-light enhancement on both infrared and visible light images, with the enhanced images guiding the training of the fusion network as part of the loss function. After training, the Edge-MobileOne Block is optimized into the structure and parameters of MobileNetV1 using structural re-parametrization techniques, further enhancing the inference speed of the fusion network. Experiments conducted on multiple public datasets have validated the superiority of the proposed method compared to 10 mainstream existing fusion methods. Even in low-light scenarios where visible light images are severely degraded, the fusion network proposed

 doi:10.20944/preprints202406.2023.v1

in this paper can generate fusion images with prominent targets and rich background textures. Moreover, the training strategy of this paper enables the fusion network to acquire low-light enhancement capabilities without additional inference computation, providing a new perspective for subsequent research in low-light scene image fusion.

Although this paper presents an image fusion method with relatively good performance, the method inevitably amplifies more noise during the process of low-light enhancement image processing, which to some extent limits the visibility of the fusion results. At the same time, subsequent visual tasks in low-light scenes, such as object detection and semantic segmentation, remain a challenge that is being faced. Therefore, the integration of image denoising technology, low-light enhancement technology, image fusion technology, and subsequent visual task models is a feasible approach. In the future, we will further improve LLE-Fuse by combining it with subsequent visual tasks, so as to ensure good visibility while also meeting the needs of machine vision tasks.

Author Contributions: Conceptualization, Song Qian, Guzailinuer Yiming and Tian Jiya; methodology, Yan Xue; software, Song Qian, Liwei Yang; validation, Song Qian, Yang Junfei; visualization, Song Qian; supervision, Guzailinuer Yiming; Writing -original draft, Song Qian; Writing -review & editing, Guzailinuer Yiming. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xinjiang Uygur Autonomous Region Tianshan Talent Programme Project, grant number 2023TCLJ02.

Acknowledgments: We thank all the editors and reviewers in advance for their valuable comments that will improve the presentation of this paper.

Data Availability Statements: The original contributions presented in the study are included in the article material, further inquiries can be directed to the corresponding author/s.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhang H, Xu H, Tian X, et al. Image fusion meets deep learning: A survey and perspective[J]. Information Fusion, 2021, 76: 323-336. [https://doi.org/10.1016/j.inffus.2021.06.008]
2. Zhang X. Benchmarking and comparing multi-exposure image fusion algorithms[J]. Information Fusion, 2021, 74: 111-131.
3. Guo C, Li C, Guo J, et al. Zero-reference deep curve estimation for low-light image enhancement[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1780-1789. [https://doi.org/10.1109/CVPR42600.2020.00185]
4. Wang D, He D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning[J]. Biosystems Engineering, 2021, 210: 271-281. [https://doi.org/10.1109/CVPR42600.2020.00185]
5. Guan D, Cao Y, Yang J, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection[J]. Information Fusion, 2019, 50: 148-157. [https://doi.org/10.1016/j.inffus.2018.11.017]
6. Jain D K, Zhao X, González-Almagro G, et al. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes[J]. Information Fusion, 2023, 95: 401-414. [https://doi.org/10.1016/j.inffus.2023.02.014]
7. Zhang Q, Zhao S, Luo Y, et al. ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2633-2642. [https://doi.org/10.1109/CVPR46437.2021.00266]
8. Chen J, Li X, Luo L, et al. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition[J]. Information Sciences, 2020, 508: 64-78. [https://doi.org/10.1016/j.ins.2019.08.066]
9. Fu Z, Wang X, Xu J, et al. Infrared and visible images fusion based on RPCA and NSCT[J]. Infrared Physics & Technology, 2016, 77: 114-123. [https://doi.org/10.1016/j.infrared.2016.05.012]
10. Li H, Wu X J, Kittler J. MDLatLRR: A novel decomposition method for infrared and visible image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4733-4746. [https://doi.org/10.1109/TIP.2020.2975984]
11. Ma J, Zhou Z, Wang B, et al. Infrared and visible image fusion based on visual saliency map and weighted least square optimization[J]. Infrared Physics & Technology, 2017, 82: 8-17.[https://doi.org/10.1016/j.infrared.2017.02.005]

12. Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623. [https://doi.org/10.1109/TIP.2018.2887342]
13. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708. [https://doi.org/10.1109/CVPR.2017.243]
14. Li H, Wu X J, Durrani T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656. [https://doi.org/10.1109/TIM.2020.3005230]
15. Li H, Wu X J, Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. Information Fusion, 2021, 73: 72-86. [https://doi.org/10.1016/j.inffus.2021.02.023]
16. Xu H, Zhang H, Ma J. Classification saliency-based rule for visible and infrared image fusion[J]. IEEE Transactions on Computational Imaging, 2021, 7: 824-836. [https://doi.org/10.1109/TCI.2021.3100986]
17. Cheng C, Xu T, Wu X J. MUFusion: A general unsupervised image fusion network based on memory unit[J]. Information Fusion, 2023, 92: 80-92. [https://doi.org/10.1016/j.inffus.2022.11.010]
18. Ma J, Tang L, Xu M, et al. STDFusionNet: An infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13. [https://doi.org/10.1109/TIM.2021.3075747]
19. Long Y, Jia H, Zhong Y, et al. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion[J]. Information Fusion, 2021, 69: 128-141. [https://doi.org/10.1016/j.inffus.2020.11.009]
20. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778. [https://doi.org/10.1007/s11042-017-4440-4]
21. Li H, Cen Y, Liu Y, et al. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion[J]. IEEE Transactions on Image Processing, 2021, 30: 4070-4083. [https://doi.org/10.1109/TIP.2021.3069339]
22. Tang L, Yuan J, Ma J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82: 28-42.[https://doi.org/10.1016/j.inffus.2021.12.004]
23. Tang L, Zhang H, Xu H, et al. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity[J]. Information Fusion, 2023: 101870.[https://doi.org/10.1016/j.inffus.2023.101870]
24. Ma J, Yu W, Liang P, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information fusion, 2019, 48: 11-26. [https://doi.org/10.1016/j.inffus.2018.09.004]
25. Ma J, Xu H, Jiang J, et al. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995. [https://doi.org/10.1109/TIP.2020.2977573]
26. Li J, Huo H, Li C, et al. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks[J]. IEEE Transactions on Multimedia, 2020, 23: 1383-1396. [https://doi.org/10.1109/TMM.2020.2997127]
27. Ma J, Zhang H, Shao Z, et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-14. [https://doi.org/10.1109/TIM.2020.3038013]
28. Liu J, Fan X, Huang Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5802-5811. [https://doi.org/10.1109/CVPR52688.2022.00571]
29. Tang L, Yuan J, Zhang H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware[J]. Information Fusion, 2022, 83: 79-92. [https://doi.org/10.1016/j.inffus.2022.03.007]
30. Liu B, Wei J, Su S, et al. Research on Task-Driven Dual-Light Image Fusion and Enhancement Method under Low Illumination[C]//2022 7th International Conference on Image, Vision and Computing (ICIVC). IEEE, 2022: 523-530. [https://doi.org/10.1109/ICIVC55077.2022.9886778]
31. Tang L, Xiang X, Zhang H, et al. DIVFusion: Darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91: 477-493. [https://doi.org/10.1016/j.inffus.2022.10.034]
32. Chang R, Zhao S, Rao Y, et al. LVIF-Net: Learning synchronous visible and infrared image fusion and enhancement under low-light conditions[J]. Infrared Physics & Technology, 2024: 105270. [https://doi.org/10.1016/j.infrared.2024.105270]
33. Reza A M. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement[J]. Journal of VLSI signal processing systems for signal, image and video technology, 2004, 38: 35-44. [https://doi.org/10.1023/B:VLSI.0000028532.53893.82]

34. Vasu P K A, Gabriel J, Zhu J, et al. Mobileone: An improved one millisecond mobile backbone[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7907-7917. [https://doi.org/10.1109/CVPR52729.2023.00764]
35. Chen Z, Fan H, Ma M, et al. FECFusion: Infrared and visible image fusion network based on fast edge convolution[J]. Mathematical Biosciences and Engineering: MBE, 2023, 20(9): 16060-16082. [https://doi.org/10.3934/mbe.2023717]
36. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141. [https://doi.org/10.1109/TPAMI.2019.2913372]
37. Jia X, Zhu C, Li M, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3496-3504. [https://doi.org/10.1109/ICCVW54120.2021.00389]
38. Toet A. The TNO multiband image data collection[J]. Data in brief, 2017, 15: 249-251. [https://doi.org/10.1016/j.dib.2017.09.038]
39. Xu H, Gong M, Tian X, et al. CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition[J]. Computer Vision and Image Understanding, 2022, 218: 103407. [https://doi.org/10.1016/j.cviu.2022.103407]
40. Zhang Y, Liu Y, Sun P, et al. IFCNN: A general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118. [https://doi.org/10.1016/j.inffus.2019.07.011]
41. Zhang H, Xu H, Xiao Y, et al. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12797-12804. [https://doi.org/10.1609/aaai.v34i07.6975]
42. Zhang H, Ma J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion[J]. International Journal of Computer Vision, 2021, 129: 2761-2785. [https://doi.org/10.1007/s11263-021-01501-8]
43. Xue W, Wang A, Zhao L. FLFuse-Net: A fast and lightweight infrared and visible image fusion network via feature flow and edge compensation for salient information[J]. Infrared Physics & Technology, 2022, 127: 104383. [https://doi.org/10.1016/j.infrared.2022.104383]
44. Xu H, Ma J, Jiang J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1):502-518. [https://doi.org/10.1109/TPAMI.2020.3012548]
45. Wang D, Liu J, Fan X, et al. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration[J]. arxiv preprint arxiv:2205.11876, 2022. [https://www.arxiv.org/abs/2205.11876v1]
46. Wang Z, Shao W, Chen Y, et al. Infrared and visible image fusion via interactive compensatory attention adversarial learning[J]. IEEE Transactions on Multimedia, 2022. [https://doi.org/10.1109/TMM.2022.3228685]
47. Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arxiv preprint arxiv:1704.04861, 2017. [https://arxiv.org/abs/1704.04861]
48. Zhu Z, Wei H, Hu G, et al. A novel fast single image dehazing algorithm based on artificial multiexposure image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-23. [https://doi.org/10.1109/TIM.2020.3024335]
49. Liu Y, Qi Z, Cheng J, et al. Rethinking the Effectiveness of Objective Evaluation Metrics in Multi-focus Image Fusion: A Statistic-based Approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 1-14. [https://doi.org/10.1109/TPAMI.2024.3367905]