Article

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

Song Qian , Guzailinuer Yiming , Ping Li , Junfei Yang , Yan Xue , Shuping Zhang *

Posted Date: 13 December 2024

doi: 10.20944/preprints202406.2023.v2

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], Guzailinuer Yiming [1], Ping Li [1], Junfei Yang [1], Yan Xue [1] and Shuping Zhang [1,*]

[1] Faculty of Information Engineering, Xinjiang Institute of Technology, Aksu, 843100, China
* Correspondence: 2015166@xjit.edu.cn

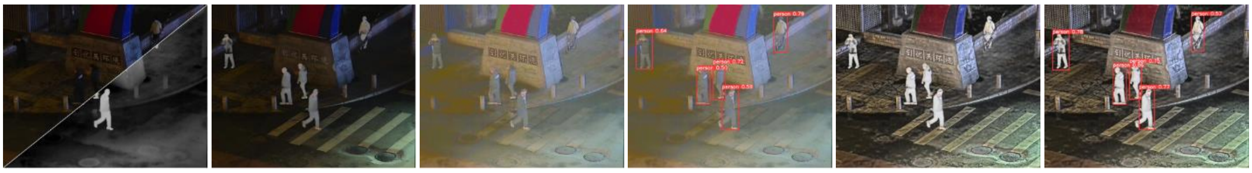
Abstract: Infrared and visible light image fusion technology integrates feature information from two different modalities into a fused image to obtain more comprehensive information. However, in low-light scenarios, the illumination degradation of visible light images makes it difficult for existing fusion methods to extract texture detail information from the scene. At this time, relying solely on the target saliency information provided by infrared images 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. The method is based on the improvement of the MobileOne Block, using the Edge-MobileOne Block embedded with the Sobel operator to perform feature extraction and downsampling on the source images. The intermediate features at different scales obtained are then fused by a cross-modal attention fusion module. In addition, the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm is used for image enhancement of both infrared and visible light images, guiding the network model to learn low-light enhancement capabilities through enhancement loss. Upon completion of network training, the Edge-MobileOne Block is optimized into a direct connection structure similar to MobileNetV1 through structural reparameterization, effectively reducing computational resource consumption. Finally, after extensive experimental comparisons, our method achieved improvements of 4.6%, 40.5%, 156.9%, 9.2%, and 98.6% in the evaluation metrics Standard Deviation(SD), Visual Information Fidelity(VIF), Entropy(EN), and Spatial Frequency(SF), respectively, compared to the best results of the compared algorithms, while only being 1.5ms/it slower in computation speed than the fastest method.

Keywords: infrared images; image fusion; low-light enhancement; feature extraction; computational resource optimization

1. Introduction

The distinct working principles of various imaging modalities lead to unique characteristics in the scenes they capture [1]. Visible light images capture light reflected from objects, offering rich texture and structural details, alongside an intuitive visual experience. However, their reliance on ambient lighting makes them vulnerable to environmental factors, such as smoke, which can compromise image quality and usability. In contrast, infrared images are generated by detecting thermal radiation emitted by objects, enabling consistent image quality across different conditions, independent of external light sources. Despite their advantages, infrared images often exhibit lower resolution and a lack of intricate background textures, limiting their effectiveness in some applications. By merging infrared and visible light images, it is possible to harness the complementary strengths of both modalities, resulting in a fused image that combines information from each source [2]. This integrated image not only retains rich scene details but also enhances target recognition and clarity. Consequently, such fusion significantly improves the comprehensiveness and accuracy of visual information, thereby increasing the practical value of the images.

Figure 1 provides an example of a low-light scenario. In Figure 1(a), the visible and infrared images captured under low illumination are presented. The result shown in Figure 1(b) stems from a contemporary image fusion technique, which, when confronted with inadequate lighting, leads to substantial degradation in the visible light image, thus lacking sufficient detail for effective fusion. Although Figure 1(c) displays the results after enhancing the low-light image with the advanced Zero-DCE (Zero-Reference Deep Curve Estimation) enhancement algorithm [3] prior to fusion, it still falls short of producing a visually satisfying outcome. In contrast, Figure 1(e) illustrates the result from the proposed algorithm, which significantly enhances visual quality. Additionally, Figures 1(d) and 1(f) demonstrate the performance of the YOLOv5s(You Only Look Once) detection algorithm [4] in Figures 1(c) and 1(e), respectively, revealing that the proposed method achieves superior pedestrian identification. An effective image fusion approach should not only aim for aesthetic quality but also ensure the preservation of essential information from the source images while highlighting critical targets. This balance is vital for facilitating higher-level visual tasks, such as target detection. By catering to the needs of both human and machine perception, image fusion technology can greatly expand its utility across various practical applications, including pedestrian re-identification [5], target detection [6], and semantic segmentation [7].



(a) VI(Visible) / IR(Infrared)

(b) ICAFusion (TMM2023)

(c) Zero-DCE + ICAFusion

(d) Zero-DCE + ICAFusion + YOLOv5s

(e) Ours

(f) Ours + YOLOv5s

Figure 1. Fusion and target detection results in the LLVIP's (Visible-infrared Paired Dataset for Low-light Vision) low-light scenario #010042 using ICAFusion (Iterative Cross-Attention Guided Feature Fusion) and the method proposed in this paper.

Recently, the integration of deep learning into image fusion has significantly transformed the field, moving away from conventional techniques like multi-scale transformation [8], subspace transformation [9], sparse representation [10], and saliency analysis [11]. These new methods typically offer superior visual outcomes. Deep learning-based fusion approaches can be classified into four primary categories: autoencoder (AE)-based, convolutional neural network (CNN)-based, generative adversarial network (GAN)-based, and unified architecture-based techniques. In the initial phase of combining deep learning with image fusion, AE-based methods relied on pre-trained encoders and decoders for extracting features and reconstructing images, applying specific strategies for fusing these features. On the other hand, CNN-based approaches execute feature extraction, fusion, and reconstruction in a seamless end-to-end process, leveraging carefully designed network architectures and customized loss functions to produce unique fusion effects. Furthermore, GAN-based methods enhance CNN techniques by adding a generative adversarial component, where the generator and discriminator engage in adversarial training, allowing the fusion outputs to approximate the source images' probability distribution in an unsupervised context.

While current deep learning-based fusion techniques are capable of effectively merging the complementary information from infrared and visible light images, several challenges remain unresolved. One key issue is that many existing methods operate on the premise that texture details primarily originate from visible light images, while salient features are sourced from infrared images. This assumption is valid under typical lighting conditions; however, in low-light or nighttime environments, the absence of sufficient illumination leads to significant degradation in visible light image quality, often obscuring crucial texture details. Consequently, these fusion methods struggle to retrieve necessary texture information, which ultimately compromises the quality of the fused output. One approach to mitigate this issue is to apply advanced low-light enhancement algorithms to improve the quality of visible light images before fusing them with infrared data. However, such enhancements can create compatibility challenges, as illustrated in Figure 1(c), where the low-light enhancement algorithm modifies the color distribution of light sources, potentially exacerbating the problem. Furthermore, using stacked network models that are tailored for different tasks may hinder the practical application and advancement of image fusion techniques. Thus, there remains a significant need to focus on designing lightweight network architectures that can address these concerns more effectively.

To address the shortcomings of current image fusion algorithms, we present a novel lightweight network for fusing infrared and visible light images, termed LLE-Fuse, specifically designed for low-light enhancement. The proposed method features a dual-branch architecture that separately extracts features from infrared and visible light images while performing downsampling at various levels. These multi-scale features are then combined through a cross-modal attention fusion module. The fused information is subsequently used to reconstruct the final image via an upsampling network. In addition, the CLAHE algorithm is employed to enhance both infrared and visible light images. The network is trained to improve its capabilities in enhancing and fusing images under low-light conditions by utilizing a combination of enhancement loss and correlation loss. To further enhance the network's representational power while keeping it lightweight, an Edge-MobileOne Block is introduced based on structural re-parameterization, functioning as both the encoder and decoder. This design aims to boost fusion performance without incurring additional computational costs during inference. The primary contributions of this paper are summarized as follows:

- This paper proposes a lightweight infrared and visible light image fusion framework based on low-light enhancement (LLE-Fuse), which can rapidly enhance the visual perception of low-light scenes, generating fused images with high contrast and clear textures.
- This paper introduces an enhancement loss designed to guide the network model to learn the intensity distribution and edge gradients of CLAHE-enhanced results, thereby enabling the model to end-to-end generate fused images with low-light enhancement effects. This effect does not incur additional computational costs during the testing phase.
- This paper designs the Edge-MobileOne Block (EMB) as the encoder and decoder of the network model, significantly improving fusion performance without increasing the computational burden during the inference phase. Additionally, this paper designs a Cross-Modality Attention Fusion Module (CMAM) that effectively integrates information from heterogeneous image modalities.

The remainder of this paper is organized as follows. In Section2, we provide a brief overview of the related work in image fusion and low-light enhancement. Section 3 delves into a detailed exposition of the proposed LLE-Fuse, including its specific modules. Section 4 presents an extensive experimental comparison demonstrating the superior performance of our method in comparison to other approaches, followed by a summary in Section 5.

2. Related Work

2.1. Deep Learning-Based Fusion Methods

Methods based on autoencoder (AE) leverage a pre-trained encoder and decoder to facilitate feature extraction and image reconstruction. These methods utilize custom-designed fusion rules tailored to the specific attributes of the source images, ensuring effective completion of the fusion process. A notable example of an AE-based fusion method is DenseFuse [12], which incorporates DenseNet [13] as its backbone and employs addition and l1-Norm as its fusion strategies. To further enhance the feature extraction capabilities of the autoencoder architecture, Li et al. [14] introduced NestFuse. This approach features a nested connection encoder and a spatial channel attention fusion module, allowing the network to capture finer detail features more effectively. Additionally, Li et al. [15] developed RFN-Nest, which utilizes residual fusion networks along with detail preservation and feature enhancement loss functions to ensure that rich texture details are maintained in the fusion outputs. Recognizing the interpretability challenges associated with fusion strategies, Xu et al. [16]

proposed a learnable fusion rule designed to assess the significance of each pixel concerning classification outcomes, thereby enhancing the interpretability of the fusion network. Furthermore, the MUFusion method introduced by Cheng et al. [17] addresses the severe degradation issues encountered during training by implementing an adaptive loss function that combines content loss with memory loss.

Methods for fusing infrared and visible light images that are based on convolutional neural networks (CNNs) utilize intricate loss functions and sophisticated network architectures to facilitate feature extraction, integration, and reconstruction. For instance, STDFusionNet [18] incorporates saliency target masks to enhance the fusion process, which helps in better preserving essential features from the images. RXDNFuse [19] merges the structural benefits of ResNet (Residual Network) [20] and DenseNet [13], enabling more comprehensive multi-scale feature extraction for effective image fusion. Li et al. [21] implemented a meta-learning strategy, allowing a single CNN model to perform image fusion across various resolutions, significantly improving the model's adaptability. Additionally, Tang et al. [22] proposed a novel semantic loss function aimed at enhancing the suitability of fused images for high-level visual tasks. In subsequent work, Tang et al. [23] revisited the integration of image fusion models with those designed for high-level visual tasks to further enhance fusion performance. Therefore, evaluating image fusion should encompass not only the visual quality and performance metrics of the results but also consider the requirements of machine vision for future high-level tasks.

Given their ability to estimate probability distributions in unsupervised learning, generative adversarial networks (GANs) have emerged as promising tools for various applications, including image fusion. FusionGAN [24] was the pioneering model to leverage GAN technology for this purpose, establishing a generative adversarial framework that enhances the textural representation of the fused images. However, relying solely on this adversarial approach can result in imbalances between the modalities in the fusion output. To tackle this issue, Ma et al. [25] introduced a dual discriminator conditional generative adversarial network (DDcGAN), which integrates both infrared and visible light images into the adversarial training process, thereby achieving a more balanced fusion outcome. Expanding on the DDcGAN concept, AttentionFGAN [26] incorporated a multi-scale attention mechanism designed to better retain foreground target information from infrared images while preserving background details from visible light images. Although training models with dual discriminators present certain complexities, Ma et al. [27] further advanced the field by proposing a GAN framework that employs multi-class constraints to effectively balance the information derived from both infrared and visible light images. Despite these advancements in visual quality, many of these methods do not fully account for how the fusion results impact subsequent high-level visual tasks. To address this gap, Liu et al. [28] developed a model that merges fusion with detection, employing a dual-layer optimization strategy. This paper aims to achieve optimal fusion results by meticulously designing a range of network architectures and loss functions, enabling CNN-based methods to implement image fusion with low-light enhancement effects.

2.2. Low-Light Scenarios-Based Fusion Methods

In scenarios where lighting conditions are inadequate, visible light images frequently undergo significant degradation during nighttime, complicating the task for current fusion methods to extract detailed texture information effectively. This often results in low-quality fused images. Research specifically addressing low-light image fusion within the image fusion domain remains sparse. For example, PIAFusion [29] acknowledges lighting conditions but relies on a simplified model that fails to adjust effectively to complex lighting environments. To counteract the detrimental effects of low light, Liu et al. [30] incorporated a visual enhancement module following the fusion network, enhancing the representation of image information. Additionally, Tang et al. [31] developed the first architecture dedicated to scene illumination fusion, which integrates low-light enhancement with dual-modality fusion to yield more informative fused images, thereby enhancing performance in subsequent visual tasks. Furthermore, Chang et al. [32] proposed an image decomposition network aimed at correcting the illumination component of the fused image, leading to improved fusion quality.

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.

2.3. Lightweight-Based Fusion Methods

Currently, research on lightweight fusion model-based studies primarily focuses on designing network models that reduce model parameters and convolutional dimension channels. IFCNN[40] achieves feature extraction and image reconstruction with only two convolutional layers in both the encoder and decoder, adjusting fusion rules based on the type of source images to implement a unified network for various fusion tasks. PMGI[41] extracts information through gradient and intensity ratio preservation, reusing and fusing features extracted from each convolutional layer. SDNet[42] addresses fusion tasks by reconstructing the generated fusion image into the squeezed network structure of the source images, forcing the fusion image to contain more information from the source images. SeAFusion[23] performs feature extraction through gradient residual dense blocks and guides the training of the fusion network using semantic segmentation task losses. FLFuse[43] completes the generation of fusion images in a lightweight and rapid manner through a weight-sharing encoder and feature exchange training strategy. However, these methods, in pursuit of lightweight design, set the network channel dimensions to be relatively shallow and employ simple fusion strategies, which fail to fully extract and fuse image features, resulting in poor performance in visual effects and performance metrics. Secondly, these overly simplistic network structures are not conducive to the training of deep learning networks, but overly complex networks are difficult to lightweight, requiring an effective method to balance the two. Structural re-parameterization technology, represented by RepVGG[48], is a new method characterized by high performance during the training phase and fast speed during the inference phase, achieving good results in multiple advanced visual tasks. However, directly using structural re-parameterization blocks designed for advanced visual tasks brings little improvement to the fusion task of infrared and visible light images. Therefore, targeted structural re-parameterization modules need to be designed for the fusion task to rapidly extract feature information from the source images.

To tackle those challenges, 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

In low-light conditions, visible light images often suffer from illumination degradation, leading to challenges for existing fusion algorithms in effectively extracting background texture details. Simply relying on infrared images to compensate for this loss proves inadequate. Consequently, the critical challenge in fusing infrared and visible light images at night lies in efficiently utilizing the information from the source images to guide the fusion network toward producing higher-quality results. One straightforward yet effective approach to address this issue is through histogram equalization. In this context, the present study employs Contrast Limited Adaptive Histogram Equalization (CLAHE) [33] for enhancing both infrared and visible light images. The network model is directed by the loss function to produce fusion images with similar characteristics, enabling it to learn low-light enhancement capabilities without adding to the inference workload.

This section aims to clarify the function of CLAHE in enhancing infrared and visible light images by showcasing visual results related to image intensity and edge details. In particular, for a registered pair of low-light infrared and visible light images, the maximum intensity and gradient outcomes derived during the computation of intensity loss and gradient loss can be expressed 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.

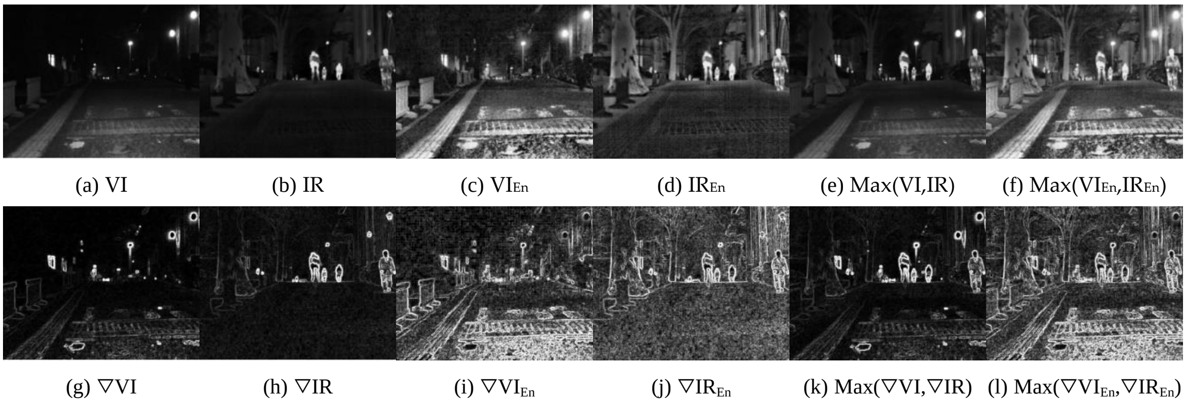
Most fusion algorithms are designed based on the assumption that visible images contain a wealth of background texture details and that they have prominent target information. This assumption holds true in most imaging environments with sufficient lighting. However, in low-light scenarios, visible images 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 better utilize the information obscured in low-light scene images, enhancing visible light images with low-light enhancement techniques is a widely adopted strategy. Among these, Contrast Limited Adaptive Histogram Equalization (CLAHE) stands out as a straightforward yet highly effective method for low-light enhancement. It significantly boosts image contrast and improves overall visual quality. Following the enhancement of both infrared and visible light images using CLAHE, the next step involves calculating the maximum intensity and maximum gradient results, which can be expressed 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 $Max\left(\cdot\right)$ signifies the selection of the maximum pixel value.



(a) VI (b) IR (c) $\mathrm{VI}_{\mathrm{En}}$ (d) $\mathrm{IR}_{\mathrm{En}}$ (e) Max(VI,IR) (f) Max($\mathrm{VI}_{\mathrm{En}}$,$\mathrm{IR}_{\mathrm{En}}$)

(g) ∇VI (h) ∇IR (i) $\nabla\mathrm{VI}_{\mathrm{En}}$ (j) $\nabla\mathrm{IR}_{\mathrm{En}}$ (k) Max(∇VI,∇IR) (l) Max($\nabla\mathrm{VI}_{\mathrm{En}}$,$\nabla\mathrm{IR}_{\mathrm{En}}$)

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

Figure 2 illustrates that both Figures 2(c) and (d) exhibit superior visual quality compared to Figures 2(a) and 2(b). This indicates that the CLAHE method effectively enhances both infrared and visible light images in low-light conditions, significantly enriching the texture detail of the scene. Consequently, this study will utilize the images in Figures 2(k) and (l) as pseudo-labels for the fusion network. By assessing the intensity distribution and edge gradients between the fusion image and the pseudo-label via an enhancement loss, the network can acquire low-light enhancement capabilities akin to those of the CLAHE method. This process enables the network to not only integrate information from both infrared and visible light images but also to better tackle fusion challenges in low-light scenarios. Furthermore, this method of guiding network training with a loss function ensures that the fusion network achieves low-light enhancement effects without incurring additional computational costs.

3.2. Network Architecture

3.2.1. Overall Network

To improve the network model's representational capability while adhering to lightweight requirements, the LLE-Fuse framework is illustrated in Figure 3. This framework effectively and efficiently carries out the image fusion task in a streamlined manner.

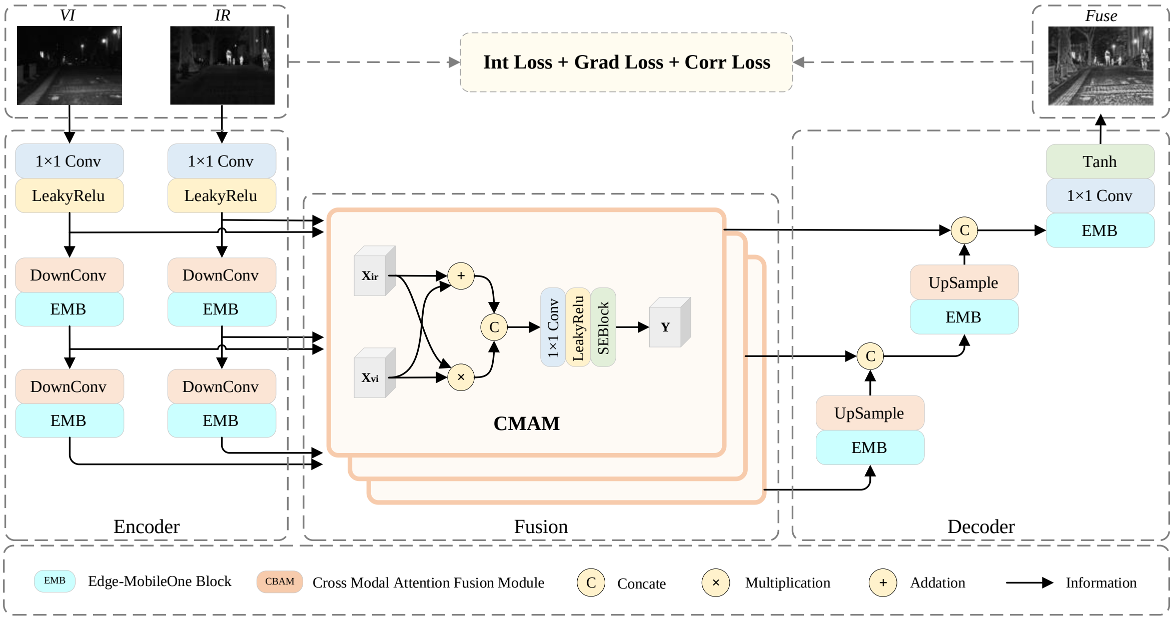


Figure 3. The overall network framework of LLE-Fuse comprises the Edge-MobileOne Block (EMB) and the Cross-Modal Attention Fusion Module (CMAM).

During the training process, the network receives infrared and visible light images through separate branches for feature extraction. The initial step involves a 1x1 convolution to capture shallow features, which are then processed in the Edge-MobileOne Block (EMB) for deeper extraction of texture and salient features. Throughout this extraction phase, the Cross-Modal Attention Module (CMAM) effectively integrates feature information from multiple stages, combining both common and distinct features while minimizing computational demands. Next, a decoder performs upsampling to merge features of varying scales, facilitating image reconstruction. Subsequently, CLAHE is applied to enhance both infrared and visible light images in low-light conditions, supporting the training of the fusion network through the loss function. Upon completing the training phase, structural re-parameterization optimizes the entire network, preparing it for the inference stage.

3.2.2. Edge-MobileOne Block

The model proposed in this paper is composed of multiple Edge-MobileOne Blocks (EMB), which are improved based on the MobileOne Block [34]. The specific structure of this module is shown in Figure 4.

In low-light scenarios, the feature network should fully extract the texture information and salient targets from infrared images to compensate for the severe degradation of visible light images.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2024 doi:10.20944/preprints202406.2023.v2

Therefore, the network should extract features at different levels, but a multi-scale feature extraction network would increase computational load. To this end, a multi-scale network structure, MobileOne Block, is introduced in the network to enrich the representational capacity of the network model during the training phase. To further reduce computational resource consumption, deep convolutional parts and point convolutions are included. Similarly, in low-light conditions, the background texture details of visible light images are lost, making the extraction of texture particularly important during the feature extraction phase. Therefore, in EMB, we replace the original structure with the Sobel operator, which enhances the expression of texture details and strengthens the network's ability to represent texture during the feature extraction phase. Concurrently, structural re-parameterization technology simplifies the multi-branch network in the training phase into a direct connection network during the inference phase, enriching the network model's representational capacity without increasing additional computational resource consumption in the inference phase.

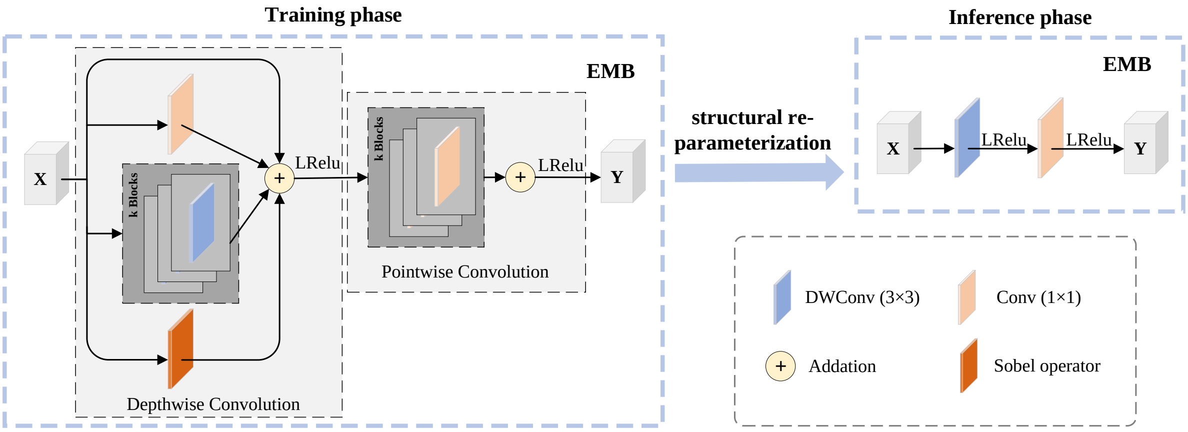


Figure 4. The network structure of the Edge-MobileOne Block (EMB).

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:

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

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

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

where, $Act\left(\cdot\right)$ 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 (Depthwise Convolution). 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

The fusion layer in our approach incorporates the Cross-Modal Attention Fusion Module (CMAM), with its specific structure depicted in Figure 3. The encoder of the fusion network extracts intermediate feature information at various scales from both infrared and visible light images. Given that these features originate from different modalities, they each emphasize distinct aspects of the scene while containing both complementary and shared information. The primary objective of the fusion module is to effectively integrate this complementary data from the distinct modalities, along with the shared information. A critical challenge lies in leveraging the unique complementary information found in one modality to effectively merge these two sets of features. For thermal targets illuminated adequately, it is essential to enhance both sets of feature information during the fusion process. However, applying methods that address complementary information similarly may risk diminishing one of the feature sets.

To effectively tackle the challenge of fusing information from diverse modalities, our approach utilizes element-wise addition to capture the complementary information present in the images from different sources. Simultaneously, element-wise multiplication is applied to extract the shared information between these modalities. The mathematical representations for both operations are as follows:

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

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

where, 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.

Using the Channel and Modality Attention Module (CMAM), the extracted common and complementary features are combined to create a feature vector, which is subsequently processed through a 1x1 convolutional layer for channel feature compression. Following this, the Channel Attention Module (SEBlock) [36] directs the network's attention toward the relevant regions, thereby improving the contrast of the identified targets. This procedure can be illustrated as follows:

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

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

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:

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

where, γ represents the weight coefficient for balancing the two losses.

The enhanced loss aims to instruct the network model in effectively learning low-light enhancement while preserving high contrast and intricate texture details. To achieve this, the model must consider both the intensity distribution and edge gradient of the fused images. This enhanced loss is composed of intensity loss and gradient loss, represented by the following formulas:

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

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

In order to enhance the model's ability to learn low-light enhancement effects, this study utilizes CLAHE to elevate the quality of both infrared and visible light images. This approach not only boosts the local contrast but also significantly minimizes noise, producing processed images with improved visual appeal, enhanced texture details, and more uniform contrast. Therefore, CLAHE will be employed to enhance both the visible light and infrared images, aiming to optimize the brightness and edge features of the fused images for superior quality in image fusion.

The intensity loss plays a crucial role in regulating the overall brightness of the fused image. To enhance the low-light effects in the fused output, this study substitutes the original infrared and visible light images with their enhanced counterparts in the common intensity loss calculation. By doing so, the maximum intensity is computed to inform the training process of the fusion network. The formula used to calculate the intensity loss is presented below:

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

where, 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:

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

where, ∇ 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:

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

where, this paper uses $corr(x, y)$ to calculate the correlation between the fused image and the source image.

4. Experiments and Analysis

This section outlines the experimental framework for training the network, covering aspects such as the dataset, experimental protocols, comparative algorithms, and evaluation criteria. Subsequently, both comparative fusion experiments and generalization tests are conducted to showcase the advantages of the proposed method. An efficiency comparison is then presented to highlight the lightweight characteristics of the approach. Finally, an ablation study is conducted to assess the effectiveness of the proposed techniques.

4.1. Experimental Setup

4.1.1. Dataset

To assess the effectiveness and generalizability of our method, comparative experiments were performed using three publicly available datasets: LLVIP [37], TNO [38], MSRS [29] and M3FD[28]. For the training phase, the LLVIP dataset, which comprises registered pairs of infrared and visible light images, was utilized. For ease of training, the image pairs from the LLVIP dataset were cropped to a resolution of 224 pixels by 224 pixels, resulting in twenty thousand pairs of images, and the CLAHE algorithm was applied to enhance the low-light conditions of these cropped images. In addition to the comparative experiments conducted on the LLVIP dataset, generalization tests were carried out using the TNO, MSRS and M3FD datasets to further validate the performance of the proposed approach.

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 comprises publicly available datasets for fusing infrared and visible light images: LLVIP, TNO, MSRS and M3FD, from which 50, 42, 30, and 50 image pairs were chosen for the algorithm comparison experiments, respectively. These experiments were carried out on a GeForce RTX 2080Ti 11GB and an Intel Core i5-12600KF, utilizing the PyTorch deep learning framework. All algorithms compared in this study were configured as specified in their respective original publications.

4.1.3. Comparative Algorithms

To further validate the superiority of the algorithm presented in this paper, LLE-Fuse was compared with 12 mainstream fusion methods, including one traditional method GTF[39], one AE-based method (DenseFuse[12]), 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 paper was primarily measured against mainstream fusion methods through both visual results and evaluation metrics.

4.1.4. Evaluation Metrics

Since fusing infrared and visible light images lacks a reference image, relying on a single evaluation metric is inadequate to demonstrate the quality of the fusion results. Thus, this paper incorporates five widely recognized image quality metrics: Standard Deviation (SD), Visual Information Fidelity (VIF), Average Gradient (AG), Entropy (EN), and Spatial Frequency (SF). Each metric offers a distinct perspective on the fusion outcomes. SD evaluates the overall contrast and distribution of the fused image. VIF measures the amount of information retained between the fused and source images by considering natural scene statistics and human visual perception. AG and SF assess image clarity through gradient and frequency analysis, respectively, while EN quantifies the informational content of the fused image. All these metrics are positively correlated; higher values indicate superior image quality. By employing these comprehensive evaluation metrics, the performance of various fusion algorithms can be assessed and compared more objectively.

4.2. Comparative Experiments

The performance of image fusion in nighttime settings is vital for this task. Consequently, the LLVIP dataset, designed specifically for urban street scenes at night, was chosen for comparative analysis. Figures 5 illustrates the visual results obtained from the proposed method alongside ten other algorithms using the LLVIP dataset. In these figures, background texture details are highlighted with green frames, while infrared salient targets are outlined in red. To enhance clarity, some images framed in solid lines have been enlarged for better detail visibility.

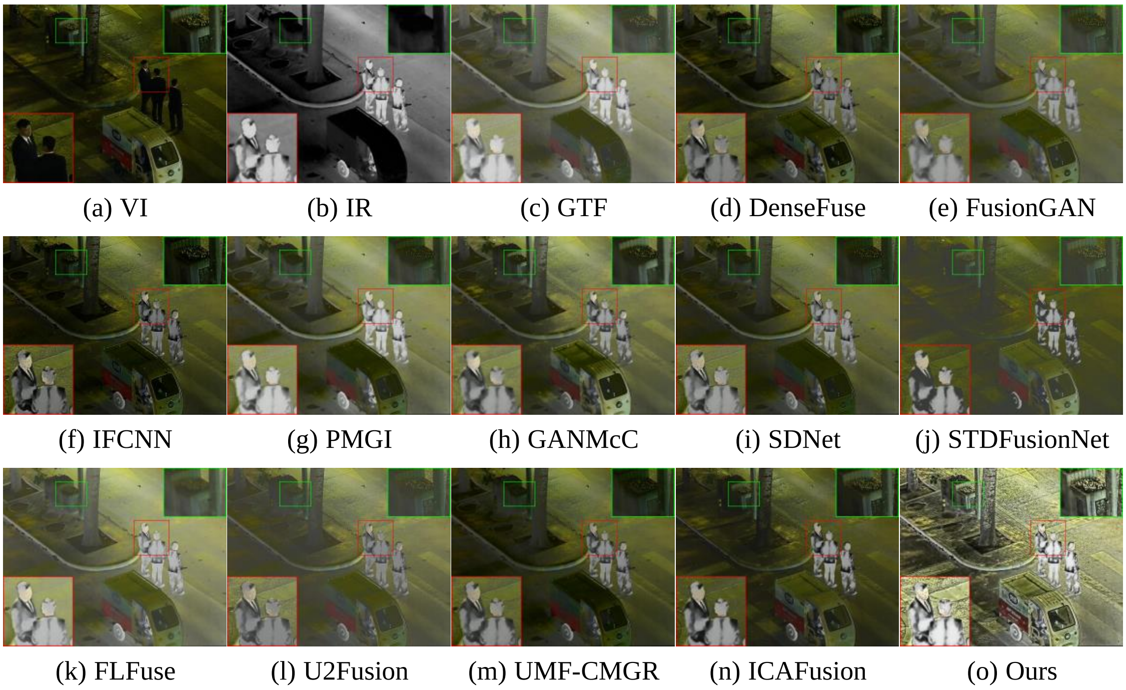

(a) VI (b) IR (c) GTF (d) DenseFuse (e) FusionGAN
(f) IFCNN (g) PMGI (h) GANMcC (i) SDNet (j) STDFusionNet
(k) FLFuse (l) U2Fusion (m) UMF-CMGR (n) ICAFusion (o) Ours

Figure 5. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the LLVIP dataset #210149.

The comparative results are depicted in Figure 5. Upon examining the background details within the enlarged green frames, it can be observed that the edge details within the green frames are nearly indiscernible in GTF, FusionGAN, and FLFuse; PMGI does render the edges visible, but the contrast between the edges and the background is low, making them easily overlooked. The remaining methods are capable of identifying the edges in the background, but they are clearly outperformed by the clarity of our method. Focusing on the enlarged red frames, only FusionGAN, PMGI, and our method have preserved the high contrast of the infrared source images, while other methods have diminished the infrared targets to varying degrees. Overall, the fused images generated by our method are the clearest, and our method is able to extract the most information from both the background and infrared targets, thus offering a significant advantage over other methods.

Table 1 presents the quantitative comparison results of our method with 12 mainstream fusion methods on the LLVIP dataset. The data in Table 1 demonstrate that our method has a significant advantage across all evaluation metrics. Specifically, our method outperformed the second-best PMGI by 4 percentage points in SD, indicating that the images generated by our method have higher contrast. In terms of VIF, our method improved by 40 percentage points compared to the best-performing DenseFuse, suggesting that the fusion results are more natural and aligned with the human visual system. Regarding the gradient assessment metric, our method showed a 156 percentage point increase compared to the top-performing IFCNN, which fully illustrates that the fusion results have extracted more gradient information. In addition, the improvements in EN and SF are also substantial, indicating that our method has a significant advantage in both visual effects and metric performance.

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

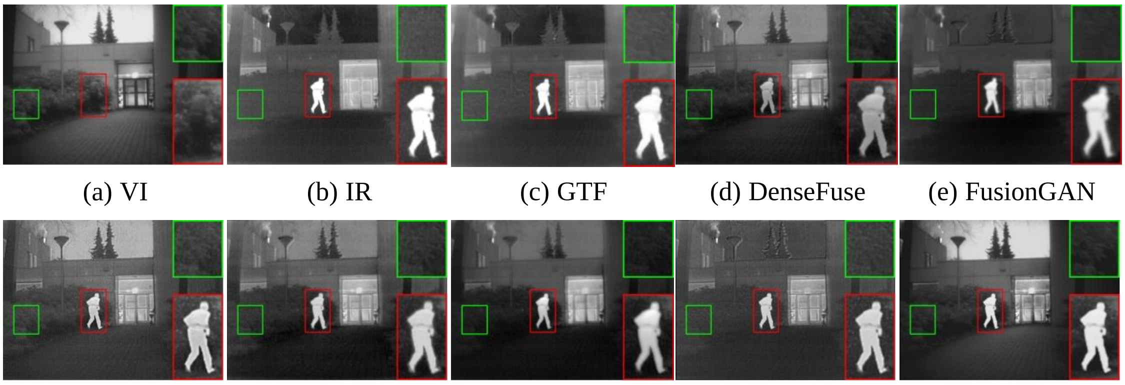
Dataset	Algorithm	SD	VIF	AG	EN	SF
LLVIP	GTF	9.0931	0.6063	2.2577	6.5562	0.0327
	DenseFuse	9.2490	0.8317	2.7245	6.8727	0.0363
	FusionGAN	8.3299	0.5322	1.9442	6.3078	0.0271
	IFCNN	8.6038	0.8094	4.1833	6.7336	0.0565
	PMGI	9.7091	0.7697	2.6571	6.9990	0.0332
	GANMcC	9.0244	0.7155	2.1196	6.6894	0.0267
	SDNet	8.9238	0.6537	3.4359	6.6793	0.0474
	STDFusionNet	6.2734	0.5222	2.9843	5.2143	0.0459
	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 capability of deep learning methods is also indicative of model performance. Therefore, in addition to comparative experiments on the LLVIP dataset, this paper also conducted experiments on the TNO, MSRS, and M3FD datasets. It is noteworthy that our method was trained on the LLVIP dataset and tested on the TNO, MSRS, and M3FD datasets.

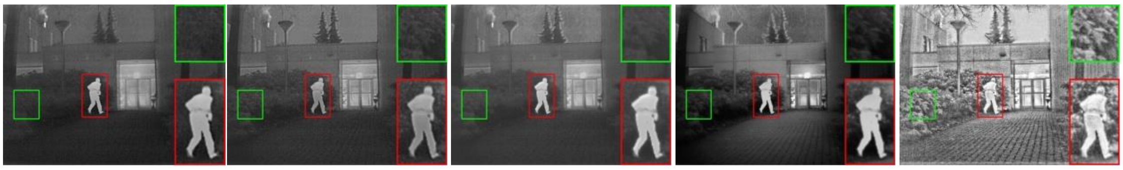
4.3.1. TNO Dataset

The TNO dataset is the most classic collection of infrared and visible light images. The visual results of the comparative algorithms and our method are shown in Figure 6. From the visual results in Figure 6, it can be observed that the fusion results of FusionGAN, GANMcC, and UMF-CMGR are relatively blurry, with a loss of scene details; DenseFuse, PMGI, SDNet, STDFusionNet, and U2Fusion retain more background details, but the scenes are still limited by the illumination degradation of visible light images, resulting in low contrast; while IFCNN, CUFD, and ICAFusion do provide better scene contrast, they still fail to offer a more visually pleasing experience. In contrast, our method demonstrates surprisingly effective results in the TNO dataset, not only providing rich background texture details and salient thermal targets but also maintaining high contrast and illuminating the entire scene.



(a) VI (b) IR (c) GTF (d) DenseFuse (e) FusionGAN

(f) IFCNN (g) PMGI (h) GANMcC (i) SDNet (j) STDFusionNet



(k) FLFuse (l) U2Fusion (m) UMF-CMGR (n) ICAFusion (o) Ours

Figure 6. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the TNO dataset #Kaptein_1123.

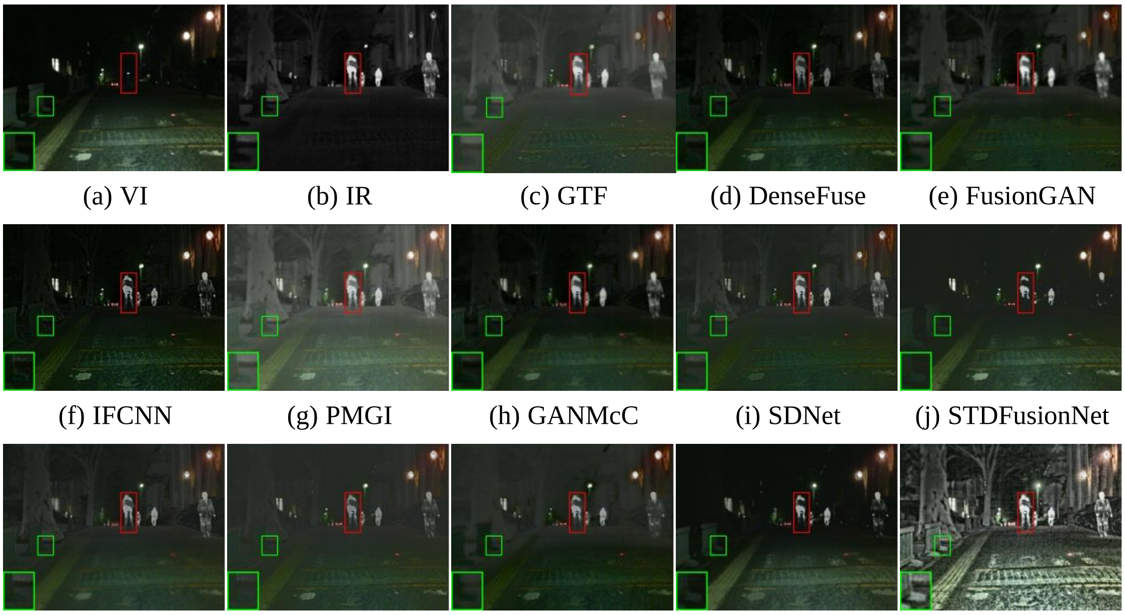
The quantitative comparison results of different methods on the TNO dataset are presented in Table 2. The data from Table 2 indicate that our method underperforms only in the VIF metric on the TNO dataset, but demonstrates significant advantages in several other evaluation metrics, an outcome attributed to the low-light enhancement capabilities of our method. Specifically, the improvements in AG and SF, which assess gradient information, are substantial, with increases of 160 and 139 percentage points, respectively. This suggests that the EMB module and enhancement loss of our method play a crucial role in extracting texture details. Furthermore, the improvements in SD and EN, with increases of 7 and 2 percentage points compared to the best-performing methods, indicate a certain advantage in the comprehensiveness of image information extraction.

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	GTF	9.4788	0.7439	3.4941	6.7632	0.0373
	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	<u>0.9746</u>	4.3846	6.9031	0.0455
	FLFuse	9.2611	0.8084	3.3691	6.3924	0.0339
	U2Fusion	8.8553	0.6787	3.4891	6.4230	0.0327
	UMF-CMGR	8.7085	0.7121	2.9727	6.5325	0.0321
	ICAFusion	9.5750	**1.0757**	4.6253	<u>7.1372</u>	0.0470
	Ours	**10.3190**	0.7849	**13.3298**	**7.3418**	**0.1218**

4.3.2. MSRS Dataset

The MSRS dataset includes some low-light scenes with insufficient illumination. This paper has selected a typical example, and the visual results of different algorithms are shown in Figure 7. Our method is capable of providing brighter scenes with prominent targets and fully exploring the background information hidden in the darkness. From the enlarged green frames, it can be seen that, aside from PMGI and our method, other methods almost fail to reveal the background details hidden in the dark. Overall, while existing mainstream fusion methods maintain the information of the source images to varying degrees, only our method provides rich texture information while maintaining the prominence of infrared targets.



(a) VI (b) IR (c) GTF (d) DenseFuse (e) FusionGAN

(f) IFCNN (g) PMGI (h) GANMcC (i) SDNet (j) STDFusionNet

(k) FLFuse (l) U2Fusion (m) UMF-CMGR (n) ICAFusion (o) Ours

Figure 7. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the MSRS dataset #01023N.

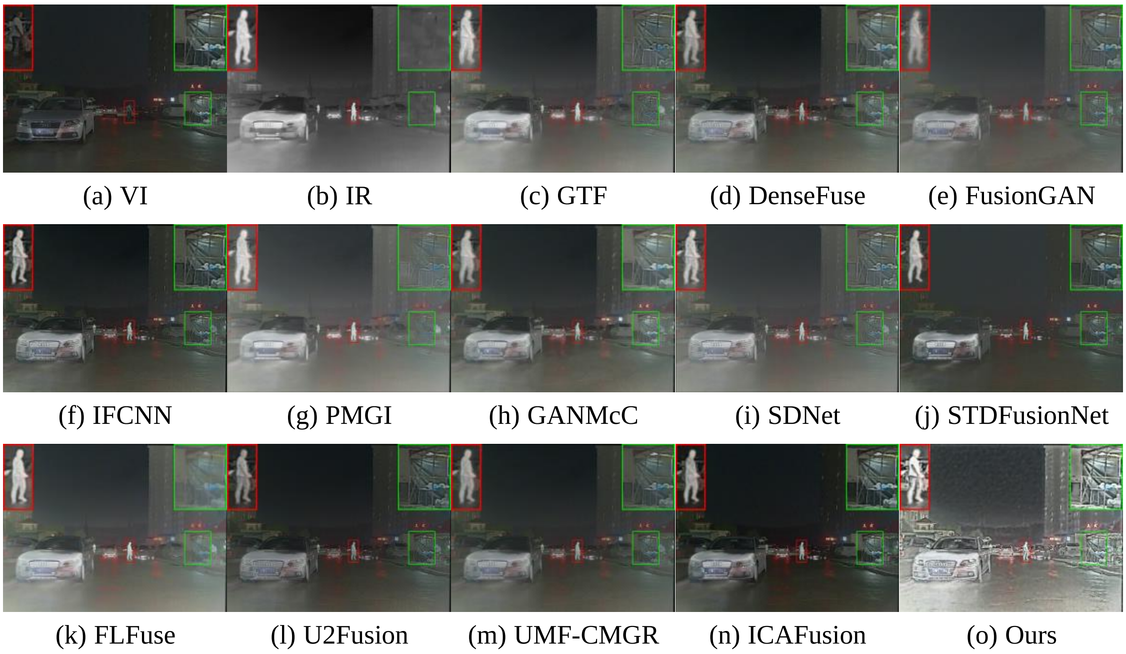
Table 3 compares LLE-Fuse's performance with other methods on the MSRS dataset across five metrics, showing LLE-Fuse's superiority in all. Its SD score reflects high contrast in fused images, while its VIF score indicates good alignment with human visual perception. LLE-Fuse's EN score also indicates rich information capture. Its AG and SF scores demonstrate effective texture preservation. Collectively, these results validate LLE-Fuse's efficacy 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	GTF	5.6669	0.4219	1.6344	5.1236	0.0218
	DenseFuse	7.5090	0.7317	2.2024	6.0225	0.0255
	FusionGAN	5.7942	0.4671	1.4470	5.4631	0.0172
	IFCNN	6.6247	0.6904	3.6574	5.8457	0.0450
	PMGI	7.5838	0.6348	2.7487	6.1807	0.0301
	GANMcC	8.0840	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
	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**

4.3.3. M3FD Dataset

In addition to the TNO and MSRS datasets, the M3FD dataset encompasses scenes with insufficient lighting, characteristic of low-light conditions. This paper presents a typical example, with the visual results from various algorithms depicted in Figure 8.


(a) VI (b) IR (c) GTF (d) DenseFuse (e) FusionGAN

(f) IFCNN (g) PMGI (h) GANMcC (i) SDNet (j) STDFusionNet

(k) FLFuse (l) U2Fusion (m) UMF-CMGR (n) ICAFusion (o) Ours

Figure 8. The comparative results of the algorithm proposed in this paper and 12 mainstream algorithms on the M3FD dataset #00621.

Upon examining the enlarged green frames, it is evident that methods such as GTF, FusionGAN, PMGI, SDNet, and FLFuse conceal background details, rendering certain background elements nearly indistinguishable in the darkness. From the enlarged red frames, it is apparent that DenseFuse, GANMcC, IFCNN, U2Fusion, and UMF-CMGR all exhibit varying degrees of attenuation in the prominence of infrared targets. Beyond this, ICAFusion demonstrates certain advantages in preserving background textures, but overall, our method significantly outperforms other approaches in depicting scene elements like "cars" and "buildings".

The quantitative comparison results on the M3FD dataset are presented in Table 4. According to the results in Table 4, our LLE-Fuse demonstrates a distinct advantage across all five evaluation metrics. The top ranking in the SD metric indicates that our method can produce fusion images with high contrast. The leading position in the VIF metric suggests that the fusion results align well with the human visual perception system. The first place in the EN metric indicates that our method can encompass rich information. The top rankings in both the AG and SF metrics indicate that our method can effectively preserve texture information from the source images in terms of gradient and frequency. In summary, the results from both qualitative and quantitative analyses substantiate the generalizability of LLE-Fuse.

Table 4. The quantitative comparison results in the M3FD dataset. The best results are marked in bold, and the second-best results are underlined.

Dataset	Algorithm	SD	VIF	AG	EN	SF
M3FD	GTF	9.5301	0.7366	2.7378	7.0239	0.0333
	DenseFuse	9.5467	0.8808	2.9959	6.9198	0.0327
	FusionGAN	8.8368	0.5800	2.2012	6.4545	0.0255
	IFCNN	9.5598	0.9026	4.8314	6.9699	0.0522
	PMGI	9.3560	0.8057	2.6775	6.8907	0.0296

GANMcC	9.7510	0.8084	2.5959	6.8776	0.0278
SDNet	9.3117	0.7517	3.5530	6.7397	0.0399
STDFusionNet	8.4282	0.8938	3.9550	6.1387	0.0470
FLFuse	9.4282	0.8087	2.0650	6.8111	0.0229
U2Fusion	9.4868	0.8224	3.4411	6.7633	0.0358
UMF-CMGR	9.5780	0.8071	2.5522	6.8253	0.0300
ICAFusion	9.1569	1.0264	3.7208	6.6036	0.0439
Ours	**10.1738**	**1.1702**	**11.5864**	**7.3626**	**0.1199**

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 5.

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. 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.

Table 5. The operational efficiency comparison results of the algorithm proposed in this paper and 12 mainstream algorithms on the LLVIP dataset.

Algorithm	**Running Time(ms)**
GTF	No data[1]
DenseFuse	267.26 ± 472.11
FusionGAN	341.21 ± 546.12
IFCNN	20.42 ± 2.13
PMGI	218.26 ± 306.99
GANMcC	726.51 ± 972.39
SDNet	111.22 ± 338.89
STDFusionNet	284.86 ± 517.61
FLFuse	**1.77 ± 14.18**
U2Fusion	511.95 ± 774.04
UMF-CMGR	260.55 ± 440.31
ICAFusion	1456.16 ± 395.66
Ours	3.27 ± 3.54

1 Since the average runtime of GTF exceeds 2 seconds, it is not included in the statistics.

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 6. According to the data in Table 6, 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 6. 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 our approach, ablation studies were conducted in this section. Experiment 1 represents our method, Experiment 2 replaces the CLAHE-enhanced images in the enhancement loss with the original source images, Experiment 3 removes the Cross-Modality Attention Module (CMAM) from the setup of Experiment 2, and Experiment 4 further removes the Edge-MobileOne Block (EMB) from the configuration of Experiment 3. The qualitative and quantitative results of the experiments are depicted in Figure 11 and Table 6, respectively.

The enhancement loss guides the network in generating fused images with low-light enhancement effects. Consequently, compared to the results of our method (Experiment 1), as shown in Figure 9(c), the scenes in the results of Experiment 2 are noticeably darker, with relatively fewer texture details. Secondly, as depicted in Figure 9(d), when the Cross-Modality Attention Module (CMAM) is removed from the basis of Experiment 2, the saliency of the fusion results is diminished. Lastly, Figure 9(e) illustrates that when the Edge-MobileOne Block (EMB) is removed from the basis of Experiment 3, the fusion network fails to effectively extract target information from the infrared image, leaning the overall scene more towards the visible light image modality.

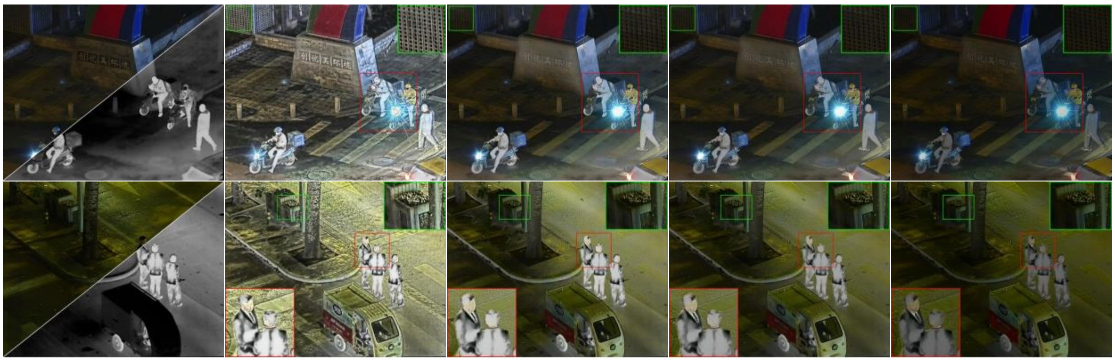

(a) VI/IR (b) Experiment1 (c) Experiment2 (d) Experiment3 (e) Experiment4

Figure 9. The ablation experimental results of the algorithm proposed in this paper on the LLVIP dataset.

Additionally, the evaluation metrics in Table 7 show varying degrees of decline in Experiments 2-4, further demonstrating the effectiveness of each module. In summary, both qualitative and quantitative results indicate that each module plays a promotional role in the overall network.

Table 7. 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>

5. Conclusions

This study introduced a novel lightweight infrared and visible light image fusion network, termed LLE-Fuse, specifically designed to address the challenges of image fusion under low-light conditions. The network employed a dual-branch architecture with Edge-MobileOne Blocks embedded with the Sobel operator for feature extraction and utilized a Cross-Modality Attention Fusion Module to integrate information from heterogeneous sources. Furthermore, this research incorporated pseudo-labels from CLAHE low-light enhancement and correlation loss into the enhanced loss function to guide the network in learning enhancement fusion capabilities in low-light scenarios. The experimental results confirmed that LLE-Fuse was capable of generating fused images with high contrast and clear textures under both low-light and normal lighting conditions, while maintaining the lightweight nature of the network. This provided an effective solution for enhancing the performance and model lightweightness of deep learning-based image fusion technology in low-light environments.

Despite the significant advancements achieved by the LLE-Fuse proposed in this study in the realm of low-light image fusion, certain limitations were noted. Specifically, the method was primarily tailored to address image fusion issues under low-illumination conditions, and it may exhibit overexposure phenomena under normal lighting conditions, leading to suboptimal fusion outcomes in brighter scenes. Additionally, if images of low-light scenes with noise are enhanced using the CLAHE method, it can amplify the image noise, thereby affecting the fusion results. This is also an issue that needs further consideration. To address this issue, future research endeavors will focus on refining the algorithm. Potential solutions include the design of network architectures capable of adapting to the varying light source distributions under both low-illumination and normal lighting conditions, thereby enabling adaptive image fusion across different lighting scenarios. Consequently, we plan to further develop an illumination classification module in the future to enhance the network's adaptability and robustness to complex environments, thereby achieving superior image fusion results under a broad range of lighting conditions.

Author Contributions: Conceptualization, Song Qian, Guzailinuer Yiming; methodology, Yan Xue; software, Song Qian, Ping Li; validation, Song Qian, Junfei Yang; visualization, Song Qian; supervision, Shuping Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was Sponsored by Xinjiang Uygur Autonomous Region Tianshan Talent Programme Project (2023TCLJ02) and Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01C349).

Data Availability Statement: Not applicable.

Acknowledgement: We are grateful to our families and friends for their unwavering understanding and encouragement.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Ethics Approval: Not applicable.

References

1. H. Zhang, H. Xu, X. Tian, et al., “Image fusion meets deep learning: A survey and perspective,” Information Fusion, vol. 76, 2021, pp. 323–336. doi: 10.1016/j.inffus.2021.06.008.
2. X. Zhang, “Benchmarking and comparing multi-exposure image fusion algorithms,” Information Fusion, vol. 74, 2021, pp. 111–131.
3. C. Guo, C. Li, J. Guo, et al., “Zero-reference deep curve estimation for low-light image enhancement,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1780–1789. doi: 10.1109/CVPR42600.2020.00185.
4. Joseph Redmon, Santosh Divvala, Ross Girshick, et al., "You only look once: Unified, real-time object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 27-30. doi: 10.1109/CVPR.2016.91.
5. D. Guan, Y. Cao, J. Yang, et al., “Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection,” Information Fusion, vol. 50, 2019, pp. 148–157. doi: 10.1016/j.inffus.2018.11.017.

6. Jiang, Chenchen, et al. "M2FNet: Multi-modal fusion network for object detection from visible and thermal infrared images," International Journal of Applied Earth Observation and Geoinformation, vol. 130, 2024, pp. 103918. doi: 10.1016/j.jag.2024.103918.
7. Q. Zhang, S. Zhao, Y. Luo, et al., “ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2633–2642. doi: 10.1109/CVPR46437.2021.00266.
8. Panda M K, Subudhi B N, Veerakumar T, et al. "Integration of bi-dimensional empirical mode decomposition with two streams deep learning network for infrared and visible image fusion," in Proc. 2022 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 493-497. doi: 10.23919/EUSIPCO55093.2022.9909631.
9. Z. Fu, X. Wang, J. Xu, et al., “Infrared and visible images fusion based on RPCA and NSCT,” Infrared Physics & Technology, vol. 77, 2016, pp. 114–123. doi: 10.1016/j.infrared.2016.05.012.
10. H. Li, X. J. Wu, J. Kittler, “MDLatLRR: A novel decomposition method for infrared and visible image fusion,” IEEE Transactions on Image Processing, vol. 29, 2020, pp. 4733–4746. doi: 10.1109/TIP.2020.2975984.
11. Panda M K, Parida P, Rout D K. "A weight induced contrast map for infrared and visible image fusion," Computers and Electrical Engineering, 2024, vol: 117, pp. 109256. doi: 10.1016/j.compeleceng.2024.109256.
12. 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, 2018. doi: 10.1109/TIP.2018.2887342.
13. G. Huang, Z. Liu, L. Van Der Maaten, et al., "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700–4708. doi: 10.1109/CVPR.2017.243.
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, 2020. doi: 10.1109/TIM.2020.3005230.
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, 2021. doi: 10.1016/j.inffus.2021.02.023.
16. H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," IEEE Trans. Comput. Imaging, vol. 7, pp. 824–836, 2021. doi: 10.1109/TCI.2021.3100986.
17. C. Cheng, T. Xu, and X. J. Wu, "MUFusion: A general unsupervised image fusion network based on memory unit," Inf. Fusion, vol. 92, pp. 80–92, 2023. doi: 10.1016/j.inffus.2022.11.010.
18. J. Ma, L. Tang, M. Xu, et al., "STDFusionNet: An infrared and visible image fusion network based on salient target detection," IEEE Trans. Instrum. Meas., vol. 70, pp. 1–13, 2021. doi: 10.1109/TIM.2021.3075747.
19. Y. Long, H. Jia, Y. Zhong, et al., "RXDNFuse: An aggregated residual dense network for infrared and visible image fusion," Inf. Fusion, vol. 69, pp. 128–141, 2021. doi: 10.1016/j.inffus.2020.11.009.
20. K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778. doi: 10.1007/s11042-017-4440-4.
21. H. Li, Y. Cen, Y. Liu, et al., "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," IEEE Trans. Image Process., vol. 30, pp. 4070–4083, 2021. doi: 10.1109/TIP.2021.3069339.
22. 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, 2022. doi: 10.1016/j.inffus.2021.12.004.
23. L. Tang, H. Zhang, H. Xu, 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," Inf. Fusion, 2023, Art. no. 101870. doi: 10.1016/j.inffus.2023.101870.
24. J. Ma, W. Yu, P. Liang, et al., "FusionGAN: A generative adversarial network for infrared and visible image fusion," Inf. Fusion, vol. 48, pp. 11–26, 2019. doi: 10.1016/j.inffus.2018.09.004.
25. J. Ma, H. Xu, J. Jiang, et al., "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," IEEE Trans. Image Process., vol. 29, pp. 4980–4995, 2020. doi: 10.1109/TIP.2020.2977573.
26. J. Li, H. Huo, C. Li, et al., "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," IEEE Trans. Multimedia, vol. 23, pp. 1383–1396, 2020. doi: 10.1109/TMM.2020.2997127.
27. J. Ma, H. Zhang, Z. Shao, et al., "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," IEEE Trans. Instrum. Meas., vol. 70, pp. 1–14, 2020. doi: 10.1109/TIM.2020.3038013.
28. J. Liu, X. Fan, Z. Huang, et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 5802–5811. doi: 10.1109/CVPR52688.2022.00571.
29. L. Tang, J. Yuan, H. Zhang, et al., "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," Inf. Fusion, vol. 83, pp. 79–92, 2022. doi: 10.1016/j.inffus.2022.03.007.

30. B. Liu, J. Wei, S. Su, et al., "Research on Task-Driven Dual-Light Image Fusion and Enhancement Method under Low Illumination," in Proc. 7th Int. Conf. Image, Vis. Comput., 2022, pp. 523–530. doi: 10.1109/ICIVC55077.2022.9886778.
31. L. Tang, X. Xiang, H. Zhang, et al., "DIVFusion: Darkness-free infrared and visible image fusion," Information Fusion, vol. 91, pp. 477–493, 2023. doi: 10.1016/j.inffus.2022.10.034.
32. R. Chang, S. Zhao, Y. Rao, et al., "LVIF-Net: Learning synchronous visible and infrared image fusion and enhancement under low-light conditions," Infrared Phys. Technol., vol. 2024, Art. no. 105270. doi: 10.1016/j.infrared.2024.105270.
33. A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," J. VLSI Signal Process. Syst. Signal, Image Video Technol., vol. 38, pp. 35–44, 2004. doi: 10.1023/B.0000028532.53893.82.
34. P. K. A. Vasu, J. Gabriel, J. Zhu, et al., "MobileOne: An improved one millisecond mobile backbone," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 7907–7917. doi: 10.1109/CVPR52729.2023.00764.
35. Z. Chen, H. Fan, M. Ma, et al., "FECFusion: Infrared and visible image fusion network based on fast edge convolution," Math. Biosci. Eng., vol. 20, no. 9, pp. 16060–16082, 2023. doi: 10.3934/mbe.2023717.
36. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141. doi: 10.1109/TPAMI.2019.2913372.
37. X. Jia, C. Zhu, M. Li, et al., "LLVIP: A visible-infrared paired dataset for low-light vision," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3496–3504. doi: 10.1109/ICCVW54120.2021.00389.
38. Toet A, "The TNO multiband image data collection, " in Proc. Data in brief, 2017, vol. 15, pp: 249-251.
39. Ma, Jiayi, et al. "Infrared and visible image fusion via gradient transfer and total variation minimization," in Proc. Information Fusion, vol. 31, 2016, pp: 100-109. doi: 10.1016/j.inffus.2016.02.001.
40. Y. Zhang, Y. Liu, P. Sun, et al., "IFCNN: A general image fusion framework based on convolutional neural network," in Proc. Information Fusion, vol. 54, pp. 99–118, 2020. doi: 10.1016/j.inffus.2019.07.011.
41. H. Zhang, H. Xu, Y. Xiao, et al., "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12797–12804. doi: 10.1609/aaai.v34i07.6975.
42. H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," Int. J. Comput. Vis., 2021, vol. 129, pp. 2761–2785. doi: 10.1007/s11263-021-01501-8.
43. W. Xue, A. Wang, and L. Zhao, "FLFuse-Net: A fast and lightweight infrared and visible image fusion network via feature flow and edge compensation for salient information," in Proc. Infrared Phys. Technol., vol. 127, Art. no. 104383, 2022. doi: 10.1016/j.infrared.2022.104383.
44. H. Xu, J. Ma, J. Jiang, et al., "U2Fusion: A unified unsupervised image fusion network," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 502–518, 2022. doi: 10.1109/TPAMI.2020.3012548.
45. D. Wang, J. Liu, X. Fan, et al., "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," arXiv preprint arXiv:2205.11876, 2022. [Online]. Available: https://www.arxiv.org/abs/2205.11876v1.
46. Z. Wang, W. Shao, Y. Chen, et al., "Infrared and visible image fusion via interactive compensatory attention adversarial learning," IEEE Trans. Multimedia, 2022. doi: 10.1109/TMM.2022.3228685.
47. A. G. Howard, M. Zhu, B. Chen, et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017. [Online]. Available: https://arxiv.org/abs/1704.04861.
48. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. "Repvgg: Making vgg-style convnets great again," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13733-13742. doi: 10.1109/CVPR46437.2021.01352.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.