electronics

MDPI

Article

BMFusion: Bridging the Gap Between Dark and Bright in Infrared-Visible Imaging Fusion

Chengwen Liu, Bin Liao and Zhuoyue Chang *

School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China
* Correspondence: 120222201314@ncepu.edu.cn

Abstract: The fusion of infrared and visible light images is a crucial technology for enhancing visual perception in complex environments. It plays a pivotal role in improving visual perception and subsequent performance in advanced visual tasks. However, due to the significant degradation of visible light image quality in low-light or nighttime scenes, most existing fusion methods often struggle to obtain sufficient texture details and salient features when processing such scenes. This can lead to a decrease in fusion quality. To address this issue, this article proposes a new image fusion method called BMFusion. Its aim is to significantly improve the quality of fused images in low-light or nighttime scenes and generate high-quality fused images around the clock. This article first designs a brightness attention module composed of brightness attention units. It extracts multimodal features by combining the SimAm attention mechanism with a Transformer architecture. Effective enhancement of brightness and features has been achieved, with gradual brightness attention performed during feature extraction. Secondly, a complementary fusion module was designed. This module deeply fuses infrared and visible light features to ensure the complementarity and enhancement of each modal feature during the fusion process, minimizing information loss to the greatest extent possible. In addition, a feature reconstruction network combining CLIP-guided semantic vectors and neighborhood attention enhancement was proposed in the feature reconstruction stage. It uses the KAN module to perform channel adaptive optimization on the reconstruction process, ensuring semantic consistency and detail integrity of the fused image during the reconstruction phase. The experimental results on a large number of public datasets demonstrate that the BMFusion method can generate fusion images with higher visual quality and richer details in night and low-light environments compared with various existing state-of-the-art (SOTA) algorithms. At the same time, the fusion image can significantly improve the performance of advanced visual tasks. This shows the great potential and application prospect of this method in the field of multimodal image fusion.

Keywords: image fusion; brightness attention unit; cross-modal information enhancement; semantic perception guidance

check for updates

Citation: Liu, C.; Liao, B.; Chang, Z. BMFusion: Bridging the Gap Between Dark and Bright in Infrared-Visible Imaging Fusion. *Electronics* **2024**, *13*, 5005. https://doi.org/10.3390/electronics13245005

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 25 November 2024
Revised: 13 December 2024
Accepted: 17 December 2024
Published: 19 December 2024

1. Introduction

Due to technical limitations and the diversity of shooting environments, it is often difficult to fully describe complex scenes from images captured by a single camera device [1,2]. Therefore, image fusion technology has become a key tool to solve this problem. Image fusion generates fused images that provide a more comprehensive description of the scene by combining information from multiple source images. Infrared and visible light fusion is of great importance in the field of image fusion [3]. Visible light images are known for their rich texture information and excellent visual adaptation. Infrared images, on the other hand, are good at capturing thermal radiation information and can highlight important targets, such as vehicles and pedestrians, in complex lighting or harsh environments. However, infrared images may not provide enough detailed information due to their single-band nature. Visible light images can also be limited in their performance under special circumstances, such as at night or when the atmosphere is heavily polluted. By fusing infrared and visible images, the limitations of the respective images can be overcome, making the final fused image more informative and recognizable. Infrared and visible light fusion technology is widely used in many fields, providing more comprehensive and accurate information for various application scenarios.

The development of infrared and visible light image fusion techniques has attracted widespread attention, especially in application areas such as military surveillance [4], target detection [5], and vehicle navigation [6]. To date, the proposed methods for IR and visible image fusion can be broadly classified into two categories: traditional methods and deep learning-based methods. Traditional methods rely on specific mathematical transformations to extract and fuse features in the source image. These include multi-scale transforms [7–10], sparse representations [11–13], significance analysis [3,14,15], subspace analysis [5,16], mixture models [17,18], and so on. However, traditional image fusion methods often require manually designed feature extraction [19–22]. This may lead to insufficient or inaccurate feature extraction when dealing with complex and variable scenes, limiting the fusion effect [23–25].

In contrast, deep learning-based methods use a data-driven approach for feature extraction and fusion. These methods utilize complex neural network models to learn and extract valuable information from the source images. The feature representation is better adapted to different scenarios through end-to-end learning and automatic learning. Deep learning methods have an excellent ability to model non-linear relationships and data-driven learning. It makes its performance in the image fusion task more flexible and has strong generalization ability, which brings significant advantages for improving the image fusion effect. Deep learning-based fusion methods are mainly classified into strategies based on CNN [26–34], GAN [35–38], and AE [28,39–43]. Among them, CNN methods achieve effective fusion through their excellent feature-processing capabilities. GAN methods enhance the realism of fused images through the generative adversarial mechanism. In addition, AE methods process feature information through the self-encoder mechanism and combine it with specific fusion strategies. This, in turn, achieves the comprehensive use of information. These different deep learning strategies together advance the development of image fusion technology. Therefore, it can show better performance in a variety of application scenarios.

Currently available algorithms perform well in many application scenarios. However, in low-light environments, these fusion methods face greater challenges. As illustrated in Figure 1, visible images tend to be under-informative in low-light conditions. While infrared images can highlight thermal targets, they are underperforming in terms of texture and structural details. Most of the existing algorithms focus only on how to fuse, without considering how to fuse in extreme environments. In low-light environments, previous algorithms ignore the degradation that occurs in visible light images at night. In the end, not only do they lose a lot of texture details in the visible light image, but they also fail to bring out the salient targets. This makes it difficult to properly obtain high-quality fused images, let alone improve the brightness and visual perception of the entire scene.

To intuitively address the problem of image fusion in low-light environments, the visible light images are preprocessed and enhanced using advanced low light enhancement algorithms [44,45]. Then, fusion techniques are applied to merge the enhanced visible and infrared images. However, processing image enhancement and fusion as independent steps often leads to incompatibility problems between them. This, in turn, affects the fusion results. This indicates the need for a more integrated approach that handles both enhancement and fusion tasks simultaneously to achieve better quality fusion results.

In order to overcome the challenges in the process of fusing infrared and visible light images, this study proposes an innovative approach that combines visual enhancement techniques with image fusion techniques. The method aims to achieve all-weather brightness enhancement alongside high-quality image fusion, ensuring that fused images are both informative and visually pleasing. It also provides a seamless transition from enhancement to fusion. This greatly simplifies the whole process and effectively solves the incompatibility problem between the two. Meanwhile, it avoids the tedious operation of frequently switching between different algorithm models in day and night scenes. This improves the generalisation ability and adaptability of the model. To this end, this study first combines Retinex theory [46], unsupervised brightness enhancement algorithm [47], and Simple Attention Module (SimAm) that can efficiently pay attention to the overall brightness features of the image [43]. A Brightness Attention Unit (BAU) module is designed for brightness adjustment. Meanwhile, the Mutually Reinforce Fusion (MRF) module was designed to enhance the fusion of cross-modal information, considering the essential differences between different modal information. Subsequently, semantic features are extracted from the visible image using an image perceptron and integrated into the feature reconstruction process. This makes the final fused image richer in semantic information and content understanding. Finally, satisfactory results are obtained by a fine loss function to guide the training in stages. Specifically, the contributions of this paper are as follows:

1. We propose an end-to-end converged network with brightness adjustment capable of solving the problem of fusing infrared-visible images under different brightness conditions in all-weather scenarios. The proposed BAU module is embedded in the network, allowing the network to learn more accurate brightness information. The MRF module is also embedded in the network, enabling the network to fuse information from different modalities more adequately.
2. We introduce adaptive learnable parameters to the traditional mathematical theory, addressing the problem of over-adjustment of brightness in the network during the fusion process. The BAU module is designed to achieve accurate scene brightness adjustment during fusion.
3. Recognizing that features of different modalities have distinct characteristics, we design the MRF module to utilize the unique properties of infrared and visible light information. The features of different modalities are supplemented and enhanced during the fusion process, reducing information loss and improving the quality of the fused image.
4. During the feature reconstruction phase, channel features are enhanced through specific knowledge guidance, making the reconstruction stage more adaptive and intelligent. This approach better preserves the feature advantages of multimodal fusion compared to other methods that simply use convolutional reconstruction.

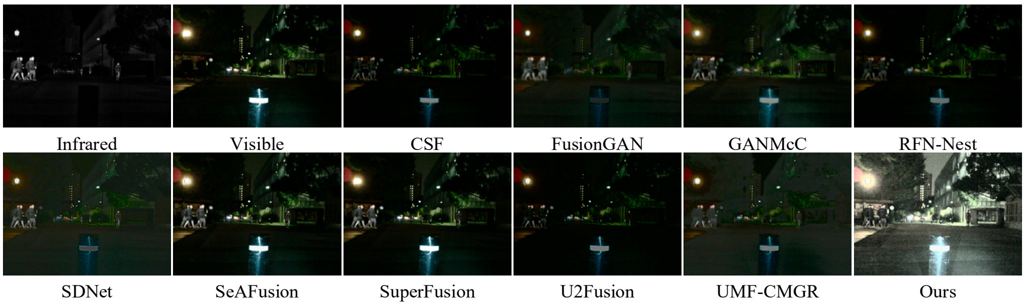


Figure 1. An example of an illumination imbalance. From left to right: IR image, visible image, fusion results of various algorithms and our proposed BMFusion. Existing methods ignore the problem of nighttime illumination degradation, leading to detail loss and thermal target degradation. Our algorithm can enhance the brightness while integrating meaningful information, mining a large amount of information lost in the dark.

The rest of the paper is organized as follows. Section 2 discusses deep learning-based methods for fusion of infrared and visible light images. It also focuses on the algorithmic study in low light scenes. The network architecture, loss function, and training details are described in Section 3. Section 4 demonstrates a comprehensive qualitative and quantitative evaluation of BMFusion, comparing it with some state-of-the-art methods. The algorithm is also validated to drive the algorithm in advanced visual tasks. Finally, conclusions are given in Section 5.

2. Related Work

In this section, existing methods for image fusion of IR and visible light are reviewed. They are CNN, GAN, and Auto-Encoder (AE)-based image fusion methods. In addition, some existing image fusion algorithms for low-light environments are focused on comparative analysis. The advantages of this paper's method over these methods are systematically discussed.

2.1. Deep Learning-Based Fusion Methods

Deep learning-based fusion methods mainly include three architectures based on CNN, GAN, and AE. The quality of fused images is improved by adaptively extracting multilevel features through the deep network structure.

CNN-based methods use the local connectivity and weight-sharing properties of convolutional layers to automatically extract and fuse information from different source images. In ICCV2017, Prabhakar [26] et al. used a CNN-based approach to solve the exposure fusion problem. However, the network is too simple to extract depth features effectively. To solve this problem, Densefuse [28] added dense blocks to the coding network to retain more useful information from the middle layer. PMGI [27] unifies the image fusion problem with texture and intensity ratio maintenance. STDfusionNet [29] reduces redundant information in fusion by significant target masks, labeling regions of human or machine interest. SeAFusion [30] proposes semantic-aware real-time fusion networks. It improves the performance of advanced vision tasks. DATFuse [31] designed an end-to-end fusion network based on a dual attention Transformer. The Dual Attention Residual Module (DARM) and Transformer Module were introduced to capture the tele-relationships. TCCFusion [32] constructed a global feature extraction branch (GFEB) using three Transformer blocks to enhance global perception. More effective long-range dependency capture was achieved. Article [33] proposed a fusion network based on progressive semantic injection and scene fidelity. It ensures that the fused features contain the complete information required for reconstruction. CDDfuse [34] proposes a relevance-driven feature decomposition fusion network. It incorporates a two-branch Transformer-CNN framework, which is better adapted to the multimodal medical image fusion task.

Generative Adversarial Networks (GANs) have received widespread attention in image fusion due to their powerful generative capabilities. GAN continuously improves the realism of generated images through adversarial loss, the game mechanism of generator, and discriminator. It effectively enhances the visual consistency of fused images. FusionGAN [35] is a pioneer in using GAN for image fusion. It establishes an adversarial game between generator and discriminator. The feasibility and advantages of GAN in image fusion are explored. However, a single adversarial game easily leads to a fusion imbalance. For this reason, the dual discriminator conditional generative adversarial network DDcGAN [36] is proposed. A more balanced fusion is obtained by engaging both IR and visible images in the adversarial game. GANMcC [37] further solves the problem of unbalanced fusion through multi-classification constraints. The balance between infrared and visible light information is improved. TarDAL [38], on the other hand, proposed a two-layer optimization network to combine fusion and detection tasks. It improves the MAP performance while enhancing the visual effect.

In image fusion, the self-encoder can effectively reconstruct the source image information and remove the noise. It also learns useful features from unlabeled data. DenseFuse [28] achieves efficient feature extraction through convolutional layers and dense blocks. However, there are limitations in remote dependency and global semantic information extraction. This leads to difficulties in capturing cross-modal associations in complex scenes. For this reason, Li et al. proposed NestFuse [39] and RFN-Nest [40]. NestFuse enhances multi-scale feature extraction through nested connections. RFN-Nest, on the other hand, introduces detail preservation and feature enhancement loss functions to learn richer features. Jian et al. [41] used an attentional mechanism to focus on salient targets and texture details. It can be implemented to cope with the redundancy problem that may be introduced by multi-scale features. DRF [42] enhances the interpretability of fusion techniques. However, it does not fully address the interpretation of fusion rules. CSF [43] proposed a learnable fusion rule by evaluating the importance of features to the classification results. Adaptive learning of fusion rules was achieved.

2.2. Image Fusion Methods in Low-Light Environments

In recent years, image fusion tasks in low-light environments have faced many challenges such as complex light degradation, detail loss, and cross-modal feature alignment. These problems make it difficult to achieve high-quality image fusion under extreme lighting conditions. In this regard, researchers have proposed a variety of solution ideas, and representative works include PIAFusion [48], DIVFusion [49], and TEXT-if [50].

PIAFusion uses layer-by-layer fusion and channel weighting mechanisms. Infrared and visible image fusion is performed by a specific feature extraction strategy. However, there are limitations in the flexibility of feature selection and capture of long-range dependencies. DIVFusion utilizes deep convolutional networks to decompose and reconstruct features layer by layer in order to capture multilevel features. However, its generalization ability is relatively weak due to the lack of explicit modeling of global features. TEXT-if proposes an innovative approach that combines textual descriptions with image feature alignment. The fusion process is guided with the help of textual semantics. However, its high dependence on input text makes it perform limitedly in purely visual scenes without textual assistance. In addition, the three methods usually rely on simple weighting or a single-feature-based approach in fusion strategy, which lacks deep enhancement of specific information. To solve the above problems, the BAU module and MRF module are designed in this paper for brightness adjustment and feature fusion, respectively. A more refined and efficient image fusion and enhancement is achieved. For the feature reconstruction stage, semantic information of visible light images is extracted by Contrastive Language-Image Pre-training (CLIP) [51].The semantic guidance is embedded into the image reconstruction process. Thus, the expression of features is continuously enhanced. At the same time, Kolmogorov-Arnold Network(KAN) [52] module is utilized for dynamic channel adjustment. The neighborhood attention enhancement mechanism is incorporated. These enable features to better adapt to environmental changes and improve reconstruction accuracy and generalization ability in different scenes.

In particular, for brightness adjustment and feature fusion, PIAFusion relies on traditional feature-level image-enhancement operations. It lacks sufficient adaptivity in the face of complex brightness variations. It is difficult to effectively handle local dark areas in low-light environments. DIVFusion uses a global contrast stretching strategy. However, it tends to introduce noise or lose local details when dealing with detail-rich regions, especially in high-contrast areas.

TEXT-if performs brightness adjustment by text description. However, it is difficult for the text to accurately express the brightness details in the image. This leads to the deficiency of its enhancement effect in regions with an obvious contrast between light and dark. In contrast, the BAU module proposed in this paper combines Retinex theory and unsupervised brightness enhancement methods to optimize brightness enhancement in low-light environments. SimAm is utilized to effectively focus on the overall brightness characteristics of the image without limiting it to local regions, resulting in a stronger global brightness adjustment capability. Transformer, on the other hand, equalizes the overall brightness by capturing long-distance dependent information, enhancing the brightness while ensuring global consistency and detail retention. The combination of SimAm and Transformer enables the BAU module to achieve more natural and effective brightness enhancement in low-light environments. The shortcomings of other methods, in detail retention and global consistency, are avoided. In addition, the method in this paper has all-weather adaptability and maintains high performance under different lighting conditions. The shortcomings of traditional methods in frequently switching or readjusting the model in day and night scenes are avoided. It also significantly improves the convenience and robustness of the model. In the MEF module, a complementary fusion strategy is adopted to reduce feature loss in information fusion.

In the feature reconstruction phase, PIAFusion, DIVFusion, and TEXT-if each have different implementations. However, all of them have obvious limitations. PIAFusion achieves decoding by simple feature channel splicing. It leads to unbalanced information fusion and is susceptible to dominant modalities, especially in complex environments. DIVFusion uses layer-by-layer decoding. Although it has advantages in local feature preservation, it does not introduce an explicit global attention mechanism in the reconstruction phase. This makes it difficult to balance the expression of details and overall structure during the reconstruction process. Especially when dealing with scenes with complex details or obvious lighting contrasts, it may cause loss of local information or unnatural effects. The semantic information in TEXT-if only plays a role in guiding the features in the early stage, and does not have a continuous role in the reconstruction stage. Simple feature splicing and decoding approaches are difficult to effectively reconstruct complete information when there is no textual description of a purely visual scene. In contrast, the method in this paper combines CLIP, KAN, and neighborhood attention enhancement. CLIP semantic vectors continuously guide feature fusion during the reconstruction process. Meanwhile, KAN adaptively weights the optimized channel features. Neighborhood attention enhancement then strengthens the complementarity of neighboring features. These strategies ensure semantic consistency, global balance, and detailed representation of the reconstructed image. It significantly outperforms other methods.

3. Methods

In this section, BMFusion is described in detail. First, the overall network architecture and network training strategy are given in Section 3.1. Then, the BAU module for brightness adjustment, the MRF module, and the progressive semantic guidance reconstruction network (PSRN) are described in Section 3.2. Finally, in Section 3.3, the loss function for training is given.

3.1. Overall Network Architecture

In this section, the network architecture used for multimodal image fusion is described in detail. As shown in Figure 2, the whole framework consists of an encoder module, MRF, and PSRN. The encoder module contains the Brightness Attention Encoder and Lossless Feature Extraction Encoder. The Brightness Attention Encoder is mainly used for brightness enhancement and feature extraction for visible light images. The Lossless Feature Extraction Encoder is used for lossless feature extraction of infrared images. MRF is responsible for fusing IR and visible light features effectively. PSRN is used to reconstruct the fused image to ensure that the information from each modality is retained.

In the training process, first, Brightness Attention Encoder and Lossless Feature Extraction Encoder are trained separately. The Brightness Attention Encoder is stacked by multiple BAUs. It gradually performs feature extraction and brightness adjustment for visible light images. The specific calculation process of Transformer and simAM in the BAU module is shown in Figure 3a and the structure of the BAUs is presented in Figure 3b. Each BAU works together with the Transformer through SimAm. It ensures fine adjustment of brightness features and capture of global information. This can enhance the naturalness and consistency of image brightness. The Lossless Feature Extraction Encoder is dedicated to IR modalities and requires only lossless feature extraction. The encoder consists of a stack of multiple Transformer blocks. It uses the same activation function and channel configuration as the Brightness Attention Encoder. This ensures that the original image information is preserved as much as possible during the feature extraction process.

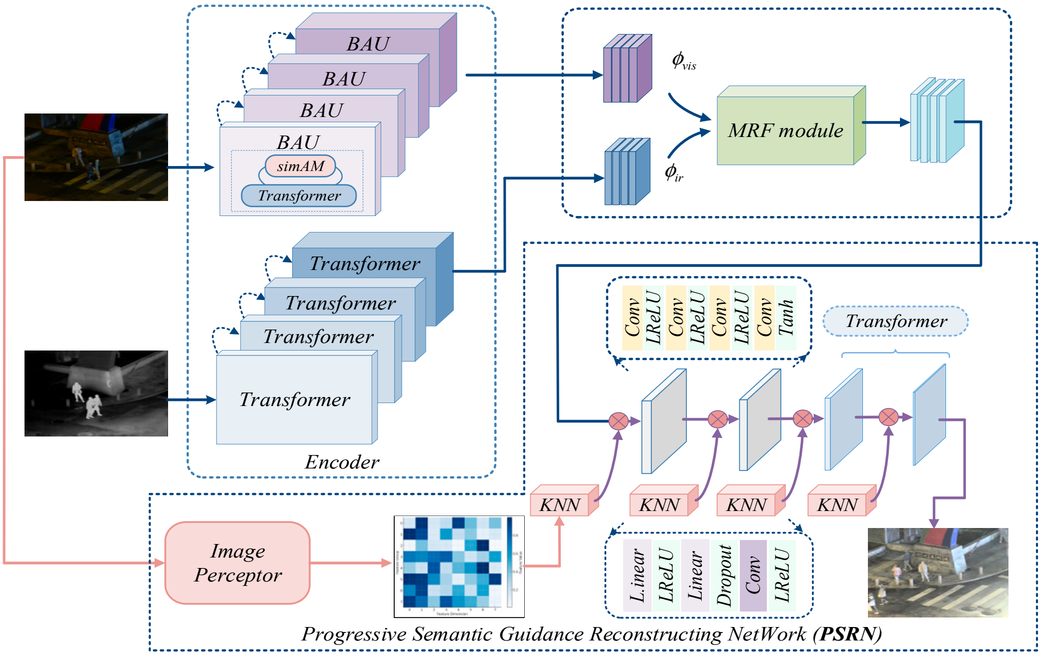


Figure 2. Overview of the BMFusion network architecture, showcasing modules for brightness adjustment, mutual feature enhancement, and progressive semantic-guided reconstruction for efficient multimodal image fusion.

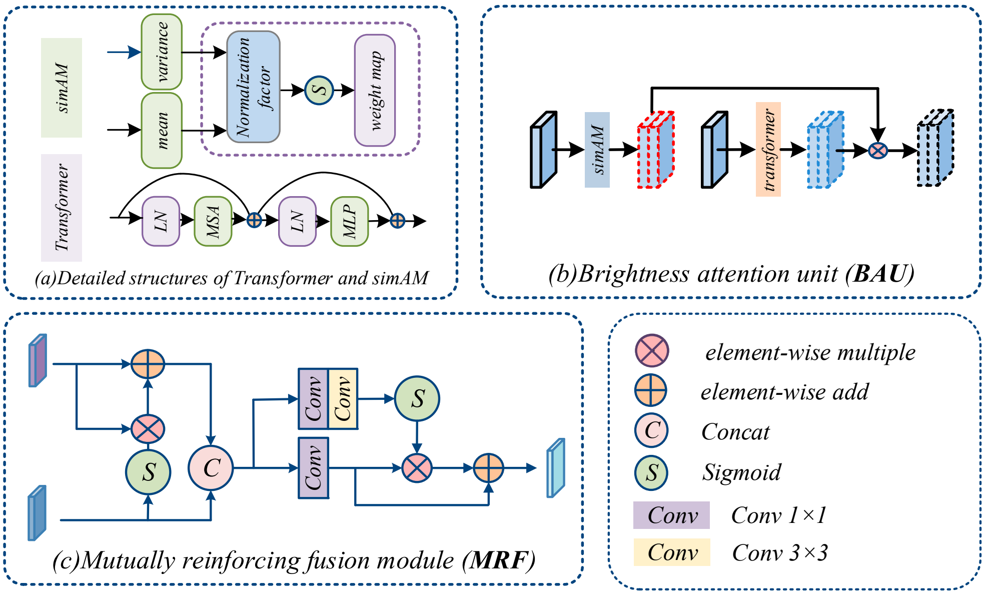


Figure 3. Illustration of detailed architectures, including Transformer, simAM mechanism, Brightness Attention Unit (BAU), and Mutually Reinforcing Fusion (MRF) modules, showcasing their roles in feature extraction, attention, and multimodal fusion.

Afterwards, the multimodal features are fused by the MRF. In this process, the features are optimized at multiple levels through alternate feature enhancement and fusion operations. It can further highlight the importance and complementarity of different modal features. The fused features have richer texture information and global semantic representation. This significantly enhances the expressive power of the fused image. These fused features are finally fed into the PSRN. It ensures that the details and semantic information of each modality are gradually integrated into the reconstruction process. Thus, high-quality image reconstruction is achieved.

Finally, in the feature reconstruction stage, PSRN is utilized to achieve finer fused image generation. The reconstruction network consists of two convolutional layers and two Transformer layers. The reconstruction of features is achieved in a layer-by-layer manner. First, the convolutional layer is used to recover the local details of the base. It ensures the accuracy of low-level features. The convolution operation efficiently processes the local features of the fused image, making the image sharper in terms of visual details. Next, the Transformer layer is used to capture the global dependencies and enhance the modeling of long-range features and the maintenance of global consistency. It makes the fused image more consistent in terms of structure and semantics. In each layer of the reconstruction phase, the global semantic vector of the visible image extracted by the image perceptron is used for semantic guidance. The image perceptron utilizes the principle of CLIP to map the semantic information of the scene in the image to an embedding space, achieving context-sensitive understanding of visual content. It ensures that high-level semantic information can be preserved when fusing features layer by layer, and improves the semantic consistency of reconstructed fused images. Specifically, the semantic information extracted by CLIP is mapped to the feature space through KAN, and weighted fusion is performed with features layer by layer. This achieves semantic enhancement for each layer of features. In addition, in order to further optimize the fused feature representation, a proximity attention mechanism is employed to match the fused features for local similarity. This enables similar feature channels to be processed more consistently in the reconstruction process. It reduces the noise effect and enhances the naturalness of the fusion result.

3.2. Important Components of the Network

3.2.1. BAU Module

The purpose of image fusion is to perceive the environment better. However, the fused image still cannot well reflect the scene information of environments with extremely low brightness. Therefore, this paper designs BAU, which combines SimAm and Transformer. It adjusts the low illumination image features while extracting features. It achieves low illumination enhancement while different modal images are fused.

According to Retinex theory, the low illumination observed image can be equal to the dot product of a clear image and illumination, i.e.,

$$y = z \otimes x \tag{1}$$

where, denotes y low-light observation image, x denotes light, and z denotes a clear image. When the input of the module is a low-light observation image and the output is a clear image, the equation can be transformed to:

$$z = y \div x \tag{2}$$

Since multiplication and division are reversible pairs of operations, the formula can be rewritten as:

$$z = y \otimes x' \tag{3}$$

Therefore, the process of restoring a low-light image to a clear image can be reduced to the process of solving x'. In this paper, BAU is embedded in the encoder part. The brightness adjustment is achieved along with feature extraction.

Define a low-light observation image extracted feature as ϕ_y Define a processed clear feature as ϕ_z. Define a light feature as ϕ_x. In the forward propagation of the network, the input of each layer is determined by the input image. In the backward propagation of the network, the output of each layer is determined by real clear image labels. Then, the features related to light intensity can be simplified to solve ϕ_x.

The brightness distribution of an image and the image itself are not independent of each other. The low brightness features are used to estimate the lighting features. In the estimation process, the previously hand-designed estimation method may have limitations and irrationality. However, if only a single convolutional kernel is used to learn the light distribution features, it can only capture local features and lacks flexibility. Therefore, in this paper, SimAm is used to replace the general convolution operation. SimAm learns the spatial information weights of input features adaptively. It achieves dynamic estimation of brightness features, avoiding the inflexibility and inaccuracy of hand-designed methods and the localization limitations of convolution kernels. It enables the brightness adjustment with higher global consistency and detail preservation. Such a design improves the model's ability to perceive light features. It is also able to enhance the brightness of low-light images more efficiently and adapt them to complex lighting environments. Then, the light features ϕ_x are transformed into a weight map generated through SimAm.

SimAm calculates the global mean and standard deviation for each channel, which are used for the adjustment of brightness characteristics:

$$\mu_c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\phi_{c,h,w} \tag{4}$$

$$\sigma_c^2 = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\phi_{c,h,w}-\mu_c)^2 \tag{5}$$

Here, $\phi_{\mathrm{c},h,w}$ denotes the value of the feature map ϕ at the Cth channel and (h,w) position. Additionally, H and W represent the height and width of the feature map.

Next, the attention weights generated by SimAm are:

$$Attention_{simam}(\phi) = \phi \times Sigmoid\left(\frac{\mu_c}{\sigma_c + \varepsilon}\right) \tag{6}$$

where ε is a very small constant used to avoid a zero denominator.

In the primary stage of feature extraction, the input visible and infrared images are subjected to preliminary convolution operations to extract their underlying features:

$$F_{vis}^{(0)} = Conv_{3\times3}(I_{vis}) \tag{7}$$

$$F_{ir}^{(0)} = Conv_{3\times3}(I_{ir}) \tag{8}$$

where I_{vis} and I_{ir} denote the input visible and infrared images, respectively, and $\mathrm{Conv}_{3\times3}$ denotes a 3 × 3 convolutional kernel for initial feature extraction for subsequent deep feature extraction.

Then, during feature extraction, the Transformer was introduced to capture remote dependencies. It processes the feature map layer by layer. First, the feature map of the previous layer is projected into the Query, Key, Value space:

$$Q = W_Q F^{(l-1)} \tag{9}$$

$$K = W_K F^{(l-1)} \tag{10}$$

$$V = W_V F^{(l-1)} \tag{11}$$

where W_Q, W_K and W_V denote the projection matrices of queries, keys, and values, respectively. F^{L-1} denotes the feature map of the previous layer. Next, the attention weights are computed:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

$$F(l) = MLP(Attention(Q,K,V)) \tag{13}$$

Finally, the SimAm-adjusted features are further processed through the Transformer to improve brightness consistency and preserve details:

$$F_{bau} = Transformer(Attention_{simam}(\phi)) \tag{14}$$

3.2.2. MRF Module

Encoder-based approaches typically rely on on-channel splicing or simple summation to fuse features from different modalities. However, such methods often overlook the inherent characteristics of each modality. Infrared images excel at capturing thermal radiation emissions, providing robust edge information, especially in low-light conditions. In contrast, visible light images offer complex spatial distributions and fine-grained background details, such as textures and colors. To address these differences and fully leverage the strengths of each modality, the MRF module is proposed in this paper. First, IR features are used to selectively enhance the critical regions in VL features, emphasizing their important details. Then, the enhanced VL features and IR features are fused to form a unified representation. Finally, a secondary enhancement process is applied to the fused features, improving edge consistency and spatial alignment. The specific architecture of the fusion module is illustrated in Figure 3c.

Visible light images usually have a more complex spatial distribution. During the fusion process, infrared features spatially replace some visible light features. This fusion rule causes some visible light information to be destroyed, resulting in the fused image having less information. To avoid this, the spatial distribution of infrared features is used to pre-enhance the visible light features before fusion. It counteracts the loss of visible light information in fusion. The process of pre-enhancement can be defined as:

$$\phi_{\mathrm{EV}} = \phi_V \oplus (\phi_V \otimes S(\phi_I)\ ) \tag{15}$$

where, ϕ_{EV} denotes the augmented visible feature. ϕ_V denotes the visible feature obtained by the encoder. ϕ_I denotes the infrared feature obtained by the encoder. S denotes the Sigmoid function, which aims at constraining the value of the infrared feature to a value between 0–1. The processed infrared features are used as weights to select the visible light features for enhancement. The enhanced visible light features can be obtained. Next, the enhanced visible features and infrared features are fused:

$$\phi_F = \mathrm{concat}(\phi_{EV}, \phi_I) \tag{16}$$

where ϕ_F denotes the fused features. *concat* $(\cdot)$ denotes the splicing on the channel. In the forward propagation process of fused features, the problem of losing edge information cannot be ignored. In order to avoid this, this paper performs a secondary enhancement of the fused features. The process of enhancement is formulated as:

$$\phi_{F\prime} = conv_1(\phi_F) \otimes (S(conv_3(conv_1(\phi_F)))) + conv_1(\phi_F) \tag{17}$$

where $\phi_{F'}$ denotes a fused feature that has been secondarily enhanced. The process of secondary enhancement can be divided into two parts: edge estimation and selective enhancement. First, a simple network structure is used for feature extraction of the fused features. This network consists of a convolution with kernel 1 and a convolution with kernel 3. This network provides a rough estimate of the spatial distribution of the fused features. The estimated spatial distribution is then processed into weights using a Sigmoid function. In the selection enhancement part, the fusion features are first spatially compressed using a convolution with a convolution kernel of 1. It forces the fused features to have tighter channel information. Next, the fusion features are spatially selected for enhancement using spatially distributed weights. The selection of weights will put most of the information while suppressing some of it, and the suppressed part may be continuously weakened until it is lost during network propagation. Information loss is not allowed in image fusion, and a short-hopping connection can perfectly avoid this problem. Specifically, the fusion features before enhancement and the fusion features after enhancement are summed up, and it avoids the loss of weakened information. The MRF module is embedded as a fusion module in the network to achieve lossless fusion of different modal features.

3.2.3. Progressive Semantic Guidance Reconstructing Network

In the feature reconstruction stage, the fused features are reconstructed layer-by-layer by two-layer convolution and a two-layer Transformer. CLIP [49] and KAN [50] modules are introduced for semantic guidance in this process.

First, the initial reconstruction of the fused features is carried out using the convolution operation:

$$F_{cnn}^{(1)} = \mathrm{Re}LU(Conv_{3\times3}(F_{fused})) \tag{18}$$

$$F_{cnn}^{(2)} = \mathrm{Re}LU(Conv_{3\times3}(F_{cnn}^{(1)})) \tag{19}$$

where, F_{fused} is the fused feature map with two layers of convolution to gradually recover the image details.

In the next two-layer Transformer module, the remote-dependent features are captured layer by layer. The feature map is projected to the query, key, and value space by the following steps:

$$Q_{cnn}^{(l)} = W_Q^{(l)} F_{cnn}^{(l-1)} \tag{20}$$

$$K_{cnn}^{(l)} = W_K^{(l)} F_{cnn}^{(l-1)} \tag{21}$$

$$V_{cnn}^{(l)} = W_V^{(l)} F_{cnn}^{(l-1)} \tag{22}$$

$$Attention_{trans}(Q_{cnn}^{(l)}, K_{cnn}^{(l)}, V_{cnn}^{(l)}) = soft\max\left(\frac{Q_{cnn}^{(l)} K_{cnn}^{(l)T}}{\sqrt{d_k}}\right) V_{cnn}^{(l)} \tag{23}$$

$$F_{trans}^{(l)} = MLP(Attention_{trans}(Q_{cnn}^{(l)}, K_{cnn}^{(l)}, V_{cnn}^{(l)})) \quad l = 3,4 \tag{24}$$

In order to enhance the semantic consistency of the reconstructed features, the semantic information of the visible images is extracted using the CLIP model:

$$F_{clip} = CLIP(I_{vis}) \in R^{1\times512} \tag{25}$$

F_{clip} denotes the 512-dimensional semantic vector extracted by CLIP. It is used to guide the reconstruction of the fused features.

Next, the CLIP features are projected by a learnable linear transformation:

$$F_{clip}^{proj} = W_{clip} F_{clip} \tag{26}$$

Semantic information is then embedded into the fusion features through element-by-element dot productions to form semantically guided features:

$$F_{kan} = F_{fused} \otimes \sigma(F_{clip}^{proj}) \tag{27}$$

where σ is the activation function. $\otimes$ denotes the element-by-element product. The semantic information is incorporated into the feature reconstruction process layer by layer through this operation.

In addition, the weighting of the fused features is adjusted by introducing a neighborhood attention mechanism to enhance the semantic similarity. First, weighting is computed for each channel of the fused features:

$$S_c = \left(\frac{F_{fused}[c] \cdot F_{clip}}{\left|\left|F_{fused}[c]\right|\right| \cdot \left|\left|F_{clip}\right|\right|}\right) \text{ for c} = 1, 2, \ldots, \mathrm{C} \tag{28}$$

$$F_{weighted}[c] = S_c \cdot F_{fused}[c] \text{ for c} = 1, 2, \ldots, \text{C} \tag{29}$$

Some of the features that are similar to the CLIP vectors are enhanced by this process.

Finally, by finding the other channels that are most similar to each channel and fusing them further:

$$KNN(c) = \text{arg}_{j \in \{1,2,\ldots,c\}, j \neq c} \text{min} \left|\left| F_{fused}[c] - F_{fused}[j] \right|\right| \tag{30}$$

$$F_{updated}[c] = \frac{1}{k} \sum_{j \in KNN(c)} F_{fused}[j] \tag{31}$$

The expression of the fused features is further refined by the fusion of neighboring features.

In PSRN, the combination of convolutional strategies and Transformer mechanisms creates a robust framework for image reconstruction by leveraging their complementary strengths. Convolutions are particularly effective at recovering fine-grained local details and ensuring textures and edges are faithfully reconstructed. Transformers, on the other hand, are adept at capturing the global context and modeling long-range dependencies, enabling the reconstruction process to account for relationships across the entire image. Together, these methods allow the reconstructed image to strike an ideal balance between localized detail and global coherence.

The reconstruction process is divided into four distinct stages, in which features at each layer are progressively refined and guided by semantic vectors derived from CLIP. These semantic vectors embed high-level contextual information, ensuring that the fused image retains a coherent and meaningful representation throughout the reconstruction. By incorporating CLIP guidance at every layer, the method ensures that global semantics are preserved without sacrificing critical local features.

A key component of this process is the KAN module, which introduces a neighborhood attention mechanism to refine spatial relationships and enhance contextual relevance. This module dynamically adapts to feature distribution at each stage, ensuring that the reconstruction process maintains spatial and semantic alignment between the different modalities. The KAN module also addresses potential conflicts between IR and VL features by selectively emphasizing complementary information and smoothing inconsistencies.

The layer-by-layer approach ensures that the reconstruction evolves in a structured and semantically consistent manner. By combining CLIP-guided semantic vectors with KAN's spatial attention, the reconstruction process is able to progress in a steadily improving trajectory. This iterative refinement ensures that the final fused image not only captures sharp local details and rich textures but also exhibits a high degree of global semantic coherence, making it suitable for both visual analysis and downstream tasks.

Overall, this multi-stage, guided framework exemplifies how the integration of convolution, Transformer modeling, semantic guidance, and neighborhood attention mechanisms can collectively drive substantial improvements in both the visual quality and semantic integrity of the reconstructed image.

3.3. Detailed Design of the Loss Function

In the task of infrared and visible fusion, a set of complex loss function combination strategies are designed in order to obtain high-quality fused images in low-light environments. These include Structure Similarity (SSIM) loss, VGG perceptual loss, gradient loss, and pixel loss. These loss functions are used for integrative constraints between different levels and types of features. It ensures the overall quality, detail clarity, perceptual effect, and semantic consistency of the image.

To measure the structural similarity between infrared and visible images, we have used SSIM loss to maintain the integrity of local structural information. The formula for SSIM is as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{32}$$

x and y denote the reconstructed fused image f and the original input image (visible image v or infrared image i) respectively, μ_x and μ_y are the mean of images x and y, σ_x^2 and σ_y^2 are the variance of images σ_y^2 and y, σ_{xy} denotes the covariance of images x and y. C_1 and C_2 are constants used to avoid the denominator being zero.

SSIM losses are defined as:

$$L_{ssim}(v,i,f) = 1 - SSIM(f,v) - SSIM(f,i) \tag{33}$$

In the fusion task, SSIM ensures that the reconstructed fused image f maintains a high structural similarity to the reference images v and i. This is particularly important for enhancing edges and details in low-light conditions.

VGG perceptual loss is used to ensure that the reconstructed images are consistent in terms of perceptual quality. This helps to maintain the semantic features of the visible and infrared images during the fusion process. Specifically, high-level features of the fused and reference images are extracted by a pre-trained VGG network. The Euclidean distance is computed on these features:

$$L_{vgg}(v,i,f) = \sum_l ||\phi_l(f) - \phi_l(v)||_2^2 + \sum_l ||\phi_l(f) - \phi_l(i)||_2^2 \tag{34}$$

where: $\phi_l(\cdot)$ denotes the feature mapping extracted through layer lth of the VGG network. Here, the features of visible v and infrared v are compared simultaneously. It ensures that the fused images are consistent in visual perception.

The introduction of VGG perceptual loss effectively addresses the high-level semantic information that may be missing from fused images. It enables semantic consistency to be maintained under low light conditions. Especially in the presence of complex lighting variations, it ensures the comprehensibility of the image.

To improve the retention of edge and texture information, gradient loss is introduced. For infrared and visible images, the gradient loss is effective in keeping them complementary in terms of edge information. It is defined as follows:

$$L_{grad}(v,i,f) = \left|\left|\nabla f - \nabla v\right|\right|_1 + \left|\left|\nabla f - \nabla i\right|\right|_1 \tag{35}$$

where ∇ denotes the gradient computed by Sobel's algorithm. The gradient loss conforms the edge characteristics of the reconstructed image to the original input image by constraining it. This improves the clarity and detail representation of the edges after image fusion.

Pixel loss is used as a direct measure of the difference between the reconstructed image and the input image at the pixel level and is defined as follows:

$$L_{pix}(v,i,f) = ||f - v||_1 + ||f - i||_1 \tag{36}$$

Pixel loss by constraining the pixel space similarity between the fused image and the input visible and infrared images.

It ensures that the fused image is consistent in brightness and color and is particularly effective in enhancing the overall brightness performance in low-light scenes.

Combining the above three loss functions to cope with the various challenges in the fusion task under low-light conditions, the brightness modulation loss function L_b and the low-light brightness modulation loss function L_b^{EN} are obtained comprehensively.

$$\begin{aligned} L_b(v,i,f) = \quad & \alpha_1 L_{ssim}(v,i,f) + \alpha_2 L_{vgg}(v,i,f) + \\ & \alpha_3 L_{grad}(v,i,f) + \alpha_4 L_{pix}(v,i,f) \end{aligned} \tag{37}$$

$$\begin{aligned} L_b^{EN}(v^{en},i,f) = \quad & \mu_1 L_{ssim}(v^{EN},i,f) + \mu_2 L_{vgg}(v^{EN},i,f) + \\ & \mu_3 L_{grad}(v^{EN},i,f) + \mu_4 L_{pix}(v^{EN},i,f) \end{aligned} \tag{38}$$

where v^{EN} refers to the high brightness reference image used to provide high brightness adjustment under night conditions after an advanced low light enhancement algorithm [43]. Here, we will briefly introduce the algorithm. It establishes a cascaded lighting learning process with weight sharing by developing a new self-calibrating lighting learning framework, which can quickly, flexibly, and robustly brighten images in low light scenes and various complex scenes in the real world. Through this algorithm, we can obtain high brightness labels. This article utilizes prior knowledge of this type, combined with the proposed BAU module, to achieve gradual brightness adjustment. α_i and μ_i (i $= 1, 2, 3, 4$) are hyperparameters that measure the importance of individual losses.

The final integrated fusion loss function $L(v, i, f)$ is:

$$L(v,i,f) = \theta \cdot L_b(v,i,f) + (1-\theta)L_b^{EN}(v^{EN},i,f) \tag{39}$$

where θ is a parameter used for the dynamic adjustment of the day and night scenes, taking $\theta = 1$ when the input image is a daytime scene and $\theta = 0$ for a nighttime scene.

This loss function is designed with full consideration of the characteristics of infrared and visible image fusion in low-light environments. Through the combination of structural similarity, perceptual loss, gradient loss, and pixel loss, the consistency of the fused images in terms of local details, global perception, and edge features is ensured. The dynamically adjusted loss function weights are adaptive to different lighting conditions. It is able to generate high-quality fused images in both day and night environments. Meanwhile, the historical brightness adjustment loss function is introduced. It enables the model to better recover the lighting and detail information in the night scene. More comprehensive quality assurance is provided for multimodal fusion.

4. Experimental Results

In this section, we first present the experimental configuration and implementation details of network training. Then, we verify the superiority of the algorithm through comparison and generalization experiments. In addition, we not only visualize the feature maps after mutual reinforcing fusion but also conduct some ablation studies to verify the effectiveness of our design. The ablation targets include BAU and MRF as well as VGG loss and gradient loss. Finally, the potential and effectiveness of our algorithm in advanced visual tasks is demonstrated in target detection experiments.

4.1. Experimental Configuration

The LLVIP dataset is particularly suited for vision tasks in low-light environments [53]. It contains aligned infrared and visible light images captured in night road scenes. Using the LLVIP dataset [54], we conducted a series of experiments on BMFusion. Both qualitative and quantitative analyses are included to assess its performance in all aspects. To further validate the generalizability of BMFusion, we also utilized the MSRS dataset [48]. It covers scenes under different lighting conditions at night with a spatial resolution of 480×640. Fifty typical nighttime image pairs were selected from the MSRS dataset for generalization analysis. Our results are compared with nine state-of-the-art (SOTA) fusion algorithms. These include five CNN-based methods, i.e., SDnet [55], U2Fusion [56], SuperFusion [57], SeAfusion, and UMF-CMGR [58], two GAN-based methods, i.e., FusionGAN and GanMcC, and two selfencoder-based methods, i.e., RFN-Nest and CSF [43]. All the above image fusion algorithms are publicly available, and we set the same parameters as reported in the original paper.

In terms of quantitative assessment, this study uses six metrics to objectively evaluate the image fusion effect. Mutual Information (MI) [59] is used to quantify the degree of information sharing between the fused image and the source image. It reflects the effective transfer of information. Average gradient (AG) [60] evaluates the richness of texture details in an image, indicating clarity and texture information. Entropy (EN) [61] Measures the amount of information in the fused image from an information theory perspective. Standard Deviation (SD) [62] Used to statistically analyze the contrast and brightness distribution of the fused image. Visual Information Fidelity (VIF) [63] evaluates image quality from the perspective of the human visual system. Finally, the Spatial Frequency (SF) [64] metric reflects the rate of change of the image gray scale, which is closely related to the clarity and texture details of the image. A comprehensive evaluation of these metrics can fully reflect the performance of image fusion algorithms. The higher the score of the fusion algorithm on these metrics, the better its fusion performance.

4.2. Implementation Details

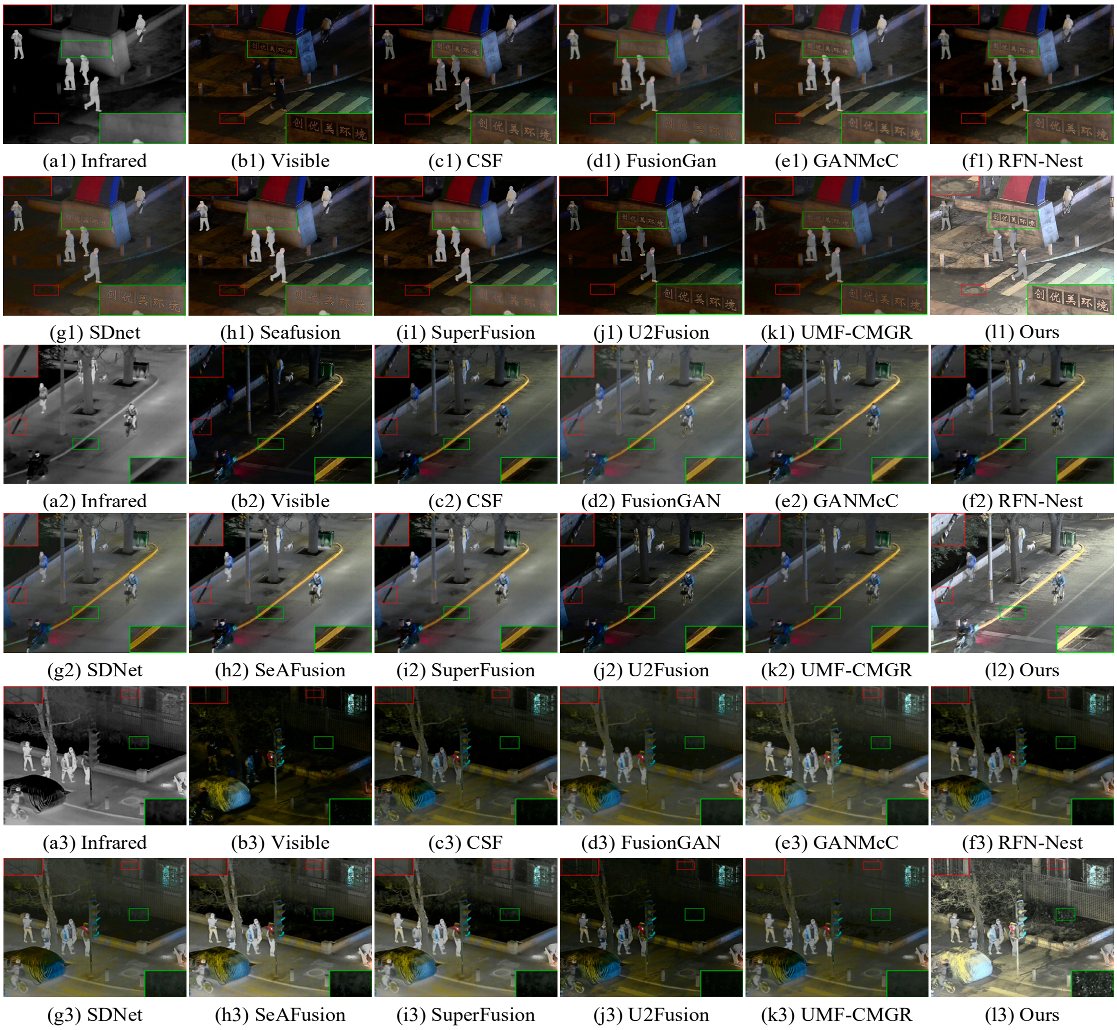
We selected the LLVIP dataset to train the fusion network proposed in this paper. Specifically, we selected 100 image pairs from the LLVIP dataset. The image pairs were then cropped into 4000 pairs of chunks of size 128 as the training set. In the training of the network, the hyperparameters were set to 10, 50, 4, and 1, respectively. (1) was set to 10, 40, 10, and 5, respectively. The batch size was set to 32, the number of training rounds was set to 30, the initial learning rate was set to 0.0001, and the training was performed using the Adam optimizer. The entire code in this paper was implemented using Pytorch 2.0. All experiments were performed on an Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz CPU and an NVIDIA P100 GPU. The Intel CPU was sourced from Intel Corporation, headquartered in Santa Clara, CA, USA. The NVIDIA GPU was sourced from NVIDIA Corporation, headquartered in Santa Clara, CA, USA.

4.3. Fusion Performance Analysis

In order to fully evaluate the performance of our method and illustrate the advantages of the method, we performed a comprehensive comparison of fusion performance with nine SOTA fusion methods on the LLVIP dataset.

4.3.1. Qualitative Results

The core objective of good nighttime low-light enhanced image fusion algorithms is to deal with illumination degradation in nighttime images. Such algorithms endeavor to extract valuable information from the original low-light image and enhance the visibility and details of the image through fusion techniques. In order to visualize the fusion performance of different algorithms on the LLVIP dataset, three pairs of infrared and visible light images were selected. The visualization results are shown in Figure 4. In the figure, we select two regions for magnification, as shown by the red and green boxes. As can be seen in Figure 4(a1–l1), none of the nine algorithms, except ours, can clearly see the outline of the manhole cover. The whole manhole cover disappears into the darkness, not to mention the texture details of the manhole cover. That is, as shown by the red box. In the green box, the fonts in FusionGan, SDnet, SeAFusio, SuperFusion, and UMF-CMGR are all very blurred and cannot be clearly rendered. The fonts in CSF, GANMcC, RFN-Nest, and U2Fusion can be rendered. However, the overall darkness makes it impossible to see the content of the fonts clearly at a glance. In contrast, our algorithm has a better overall visual experience and is relatively bright. It allows one to see the information conveyed in the image effortlessly at a glance. In Figure 4(a2–l2), the sewer covers in all the algorithmic scenes are submerged in darkness, except for our algorithm, which is shown in the green box. In the red box, it can be seen that only our algorithm can clearly see the color and texture details of the curb cables. Compared to the other algorithms, our results all contain more prominent targets and richer and clearly presented texture information. It also provides a visually brighter, high-contrast scene. Even the information about objects that have been mostly lost in the darkness can be shown clearly, more similar to a scene in daylight.



(a1) Infrared (b1) Visible (c1) CSF (d1) FusionGan (e1) GANMcC (f1) RFN-Nest

(g1) SDnet (h1) Seafusion (i1) SuperFusion (j1) U2Fusion (k1) UMF-CMGR (l1) Ours

(a2) Infrared (b2) Visible (c2) CSF (d2) FusionGAN (e2) GANMcC (f2) RFN-Nest

(g2) SDNet (h2) SeAFusion (i2) SuperFusion (j2) U2Fusion (k2) UMF-CMGR (l2) Ours

(a3) Infrared (b3) Visible (c3) CSF (d3) FusionGAN (e3) GANMcC (f3) RFN-Nest

(g3) SDNet (h3) SeAFusion (i3) SuperFusion (j3) U2Fusion (k3) UMF-CMGR (l3) Ours

Figure 4. Qualitative comparison of BMFusion with 9 state-of-the-art methods in different scenes on LLVIP datasets.

4.3.2. Quantitative Results

We further performed a quantitative comparison of 50 pairs of images from the LLVIP dataset, as demonstrated in Table 1. Our method scored first place in all five metrics and third place in the MI comparison. This achievement proves the efficiency of our method. The performance in MI illustrates the advantages of our fusion results in information sharing. Although our method is slightly inferior to SeAFusion and SuperFusion in the metric of MI. However, our method significantly outperforms the other comparison algorithms in terms of brightness reinforcement and scene clarity. This strategy of increasing the overall scene brightness, however, affects the linear correlation with the source image, thus dropping the ranking to third on MI. However, this ranking is still within acceptable limits, given the clear advantage of our method in terms of visual effect. The high score for AG shows the superiority of our method in detail information retention. The best performance in EN reflects the rich information content of the fused images. In addition, our high scores on SD and VIF further confirm the improved image contrast and optimized visual quality. The excellent performance of SF, on the other hand, demonstrates a significant improvement in image sharpness and texture details. Overall, our method achieves significant results in enhancing the overall brightness of the scene. This enables the fused images to outperform other methods in terms of clarity and visual effect. Finally, we arbitrarily selected 20 pairs of images from the LLVIP dataset and made a line graph after counting the corresponding metrics, as shown in Figure 5.

Table 1. Comparison of metrics on LLVIP. The red font indicates the best indicator, the blue font indicates the second best indicator, and the green font indicates the third best indicator.

	SD	MI	VIF	AG	EN	SF
CSF	9.0574	2.5046	0.7940	2.5753	6.6975	0.0322
FusionGAN	8.3254	2.8171	0.5319	1.9468	6.3083	0.0271
GANMcC	9.0199	2.6817	0.7152	2.1229	6.6899	0.0267
U2Fusion	8.1659	2.2748	0.7158	3.2891	6.3597	0.0433
RFN-Nest	9.2655	2.5545	0.8198	2.1579	6.8624	0.0248
SDNet	8.9246	2.9725	0.6534	3.4387	6.6800	0.0474
SeAFusion	9.4885	3.7725	0.9882	3.8317	7.2353	0.0514
SuperFusion	9.4757	4.0397	0.8982	3.1196	7.1280	0.0437
UMF-CMGR	8.0539	2.6817	0.5796	2.5041	6.4620	0.0389
Ours	10.0742	3.7399	1.3379	6.1919	7.3673	0.0814

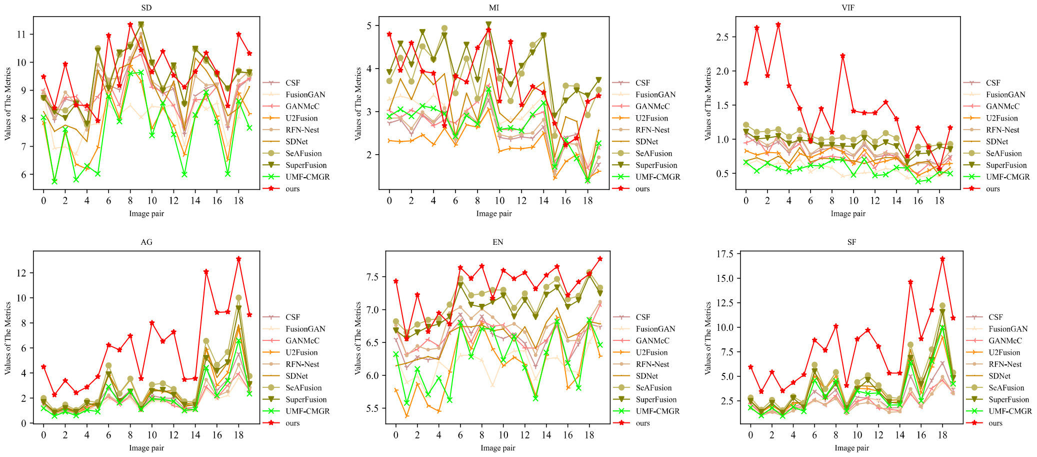


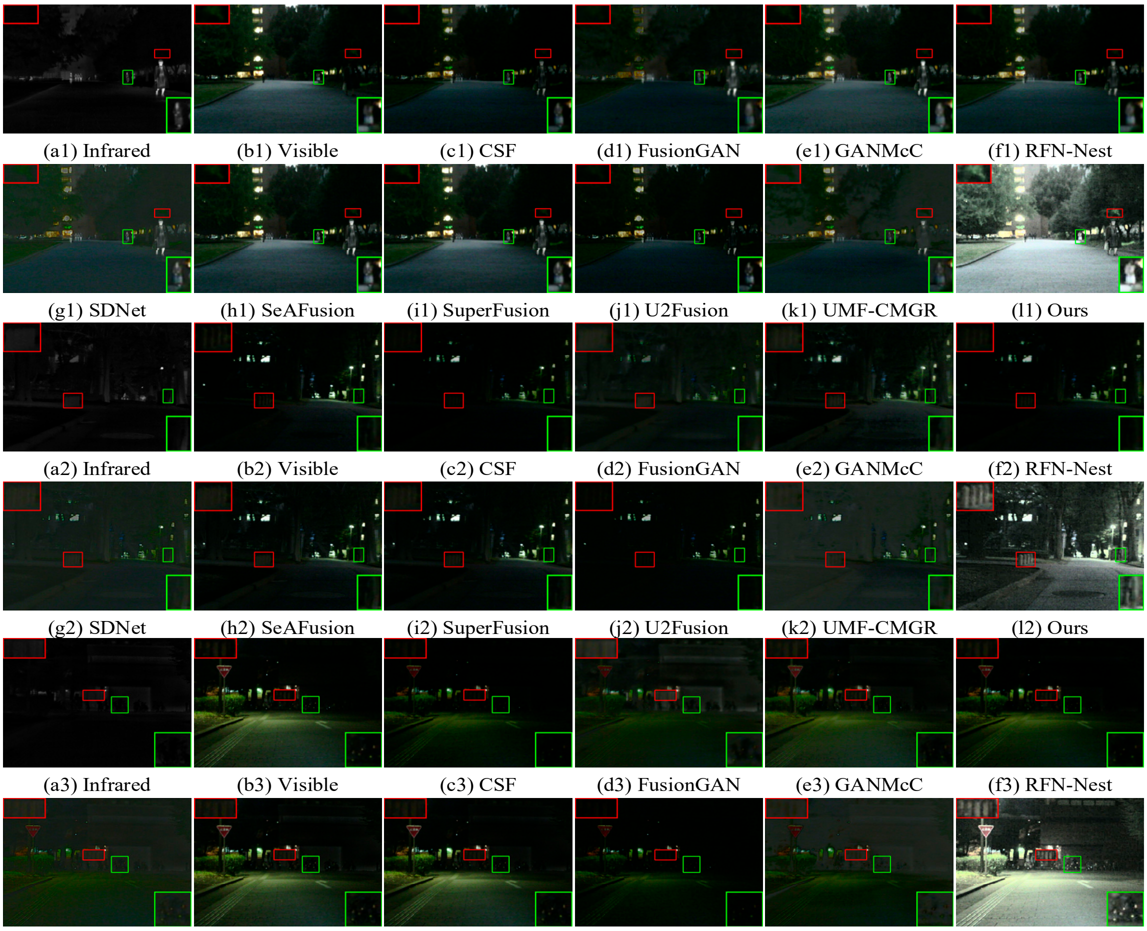
Figure 5. Quantitative results of six metrics, i.e., SD, MI, VlF, AG, EN, and SF, on any 20 image pairs from the LLVIP dataset. Nine SOTA methods are used for comparison.

4.4. Generalisation Experiments

In the field of deep learning, the generalization ability of a model is a key metric for assessing its performance. Therefore, we selected 50 pairs of nighttime infrared and visible images from the MSRS dataset. Generalization tests were performed on our BMFusion model. In particular, it is noted that the BMFusion model was trained only on the LLVIP dataset. It has not been evaluated on the MSRS dataset without any specific tuning or optimization of that dataset directly. The generalization experiments demonstrate the adaptability and stability of our model when dealing with different data sources.

4.4.1. Qualitative Results from the MSRS Dataset

The results of the qualitative comparison of the MSRS dataset are shown in Figure 6. As can be seen in Figure 6(a1–l1), our algorithm is able to better highlight pedestrians on the far side of the road. This is due to the BAU module that we have designed to improve the overall contrast of the image. As a result, there is a significant improvement in the salient targets compared to other algorithms, i.e., shown in the green box. The texture of the leaves submerged in the darkness is well revealed in the red box. This can be seen by zooming in and looking at the stone pillars at the bottom of the roadside intersecting with power lines and tree trunks in Figure 6(a2–l2) and the entrance to the building and the motorbike vehicles on the road in Figure 6(a3–l3). In very dark scenes, the fusion results of other algorithms have been completely unable to present the fusion results well. The fusion results of some algorithms even lose many texture details in the visible image while failing to highlight the significant targets. On the contrary, our method mines out a large amount of scene information hidden in the darkness. It contains both the high contrast of infrared images and the rich texture details of visible light images. Compared to the LLVIP dataset, the images in the MSRS dataset have lower brightness, contrast, and sharpness. The problem of illumination degradation in nighttime images is even more difficult to solve. This is more demanding on the performance, robustness, and applicability of our algorithm. Even so, our method has a better overall visual perception. This further illustrates the superiority of our proposed algorithm.



(a1) Infrared (b1) Visible (c1) CSF (d1) FusionGAN (e1) GANMcC (f1) RFN-Nest

(g1) SDNet (h1) SeAFusion (i1) SuperFusion (j1) U2Fusion (k1) UMF-CMGR (l1) Ours

(a2) Infrared (b2) Visible (c2) CSF (d2) FusionGAN (e2) GANMcC (f2) RFN-Nest

(g2) SDNet (h2) SeAFusion (i2) SuperFusion (j2) U2Fusion (k2) UMF-CMGR (l2) Ours

(a3) Infrared (b3) Visible (c3) CSF (d3) FusionGAN (e3) GANMcC (f3) RFN-Nest

(g3) SDNet (h3) SeAFusion (i3) SuperFusion (j3) U2Fusion (k3) UMF-CMGR (l3) Ours

Figure 6. Qualitative comparison of BMFusion with 9 state-of-the-art methods in different scenes on MSRS datasets.

4.4.2. Quantitative Results for the MSRS Dataset

We performed a quantitative comparison of 50 pairs of images from the MSRS dataset. The results of the comparison of the different algorithms for the six metrics are shown in Table 2. Our method achieved first place on all six metrics. On the MI metric, our method demonstrates excellent information-sharing ability. The high score of AG reflects the rich texture details of the image. The excellent performance of EN reveals that the fused image contains a large amount of information. In SD and VIF, our method also performs excellently. It shows significant improvement in image contrast and excellent visual perception performance. The high score of SF further proves the effectiveness of our method in improving image clarity. In summary, our method shows strong fusion capabilities both in terms of image detail retention, information content, visual quality, and clarity. This is confirmed by the performance on the MSRS dataset. Similarly, we arbitrarily selected 20 pairs of images from the MSRS dataset and made a line graph after counting the corresponding metrics, as shown in Figure 7.

Table 2. Comparison of metrics on MSRS. The red font indicates the best indicator, the blue font indicates the second best indicator, and the green font indicates the third best indicator.

	SD	MI	VIF	AG	EN	SF
CSF	5.1253	1.9904	0.5467	1.1249	4.5089	0.0173
FusionGAN	4.6643	1.7124	0.3565	0.9617	4.8367	0.0123
GANMcC	7.1819	2.2346	0.6971	1.4717	5.6018	0.0173
U2Fusion	4.2689	1.8005	0.4188	1.2306	3.9947	0.0219
RFN-Nest	5.7250	2.0711	0.6357	1.1042	5.0092	0.0156
SDNet	4.0533	1.4478	0.2737	1.4639	4.2654	0.0216
SeAFusion	6.4184	3.1782	0.9227	2.0949	5.6790	0.0293
SuperFusion	6.3411	3.0876	0.8915	1.9017	5.5546	0.0286
UMF-CMGR	4.5016	1.7318	0.2157	1.3277	4.8857	0.0191
Ours	9.4120	4.6922	1.1587	4.2865	7.2781	0.0464

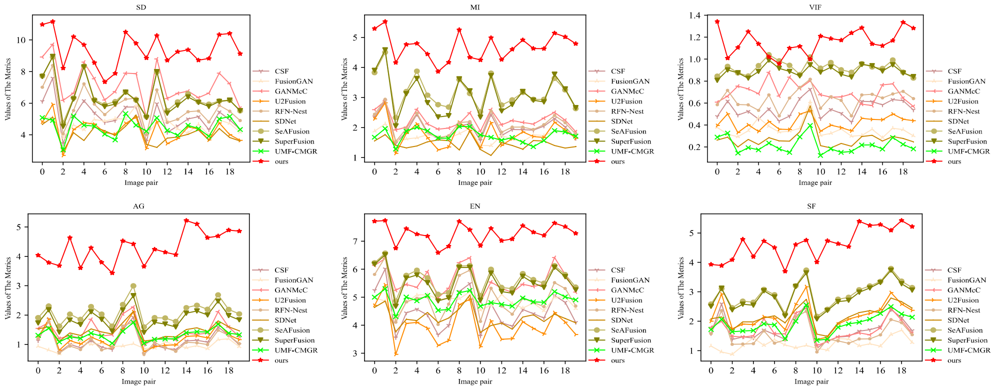


Figure 7. Quantitative results of six metrics, i.e., SD, MI, VlF, AG, EN, and SE, on any 20 image pairs from the MSRS dataset. Nine SOTA methods are used for comparison.

4.5. Efficiency Comparison

In order to provide an overall evaluation of the different algorithms, we give the average running time and model size of the different methods in Table 3 The average runtime here refers to the average time taken by the network to generate a fused image. The model size refers to the amount of memory required to store the model parameters. These parameters include weights, biases, and other necessary structural elements in the model architecture.

Table 3. Comparison of operational efficiency. The red font indicates the best indicator, the blue font indicates the second best indicator.

	Speed/s	Model-Size/MB
CSF	14.4062	4.0673
FusionGAN	0.6257	7.248
GANMcC	1.0622	10.9472
U2Fusion	0.4897	2.5878
1RFN-Nest	0.838	28.7265
SDNet	0.2041	0.7148
SeAFusion	0.1925	0.6513
SuperFusion	1.7092	22.4628
UMF-CMGR	0.344	7.2255
Ours	0.34	1.86

As can be seen, CSF indicates whether a pixel needs to be fused into the result by evaluating the contribution/significance of each pixel in the feature map. This is very time consuming. In the field of image fusion, the use of smaller deep learning models has significant advantages. First, smaller models enable faster data processing and inference, which is particularly critical for real-time image fusion applications. Second, these models have a lower demand for computational resources. This makes them suitable for running on resource-limited devices, such as mobile and embedded systems. In addition, small models are easier to deploy and maintain, reducing the risk of overfitting. In short, small models bring the dual benefits of flexibility and efficiency to image fusion tasks while maintaining reasonable performance. SeAFusion has been designed to be lightweight in terms of network design, taking into account the real-time requirements of preprocessing operations. It is the fastest algorithm for all datasets. The squeeze decomposition network designed in SDNet is also a lightweight network. Our average runtime and model size are second only to SeAFusion and SDNet. Our network combines the two tasks of brightness reinforcing and image fusion. A reasonable fusion strategy is designed according to the characteristics of different modal information itself. In order to achieve good results, it needs to consume a certain amount of time. However, our method still has comparable operation efficiency. This fully demonstrates the excellent computing efficiency and adaptability of our method.

4.6. Mutually Enhanced Fusion Visualization

In the framework proposed in this study, we carefully design the l loss function. They achieve precise control of the image reinforcing, feature extraction, feature selection, and image reconstruction processes under low-light conditions. Figure 8 shows some of the feature maps after processing by the MRF module. It is obvious from these images that the network successfully preserves the texture features of the enhanced visible image. It effectively integrates the salient target features in the infrared image into the fused feature maps. A strong validation of the efficiency of our network model in performing feature fusion.

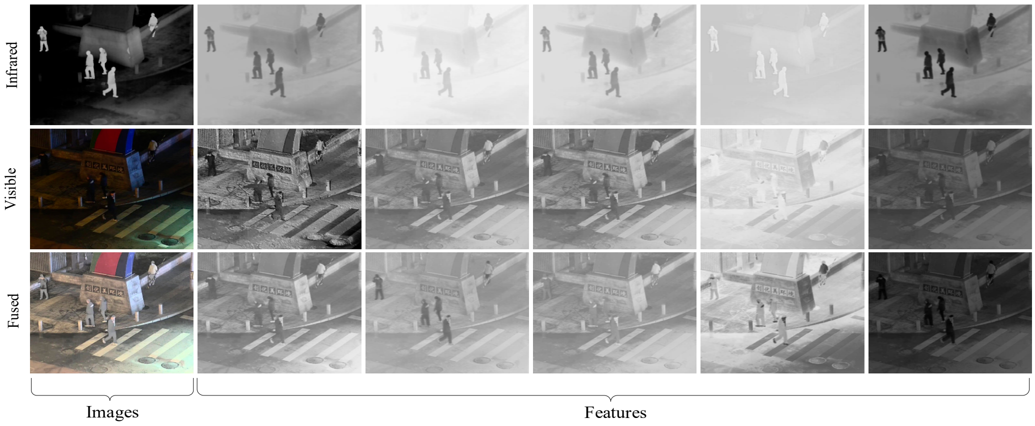


Figure 8. Visualized results of images and feature maps. The first column presents the infrared image, visible image, and fused image, respectively. The next five columns show the feature maps corresponding to the infrared, visible, and fused images in various channel dimensions.

4.7. Ablation Experiment

4.7.1. BAU and MRF Ablation Analyses

The important components of our fusion model are the brightness modulation unit module and the cross-modal mutual reinforcing fusion module. Therefore, we also performed ablation experiments on these two modules, and the experimental results are shown in Figure 9. After removing the BAU module, the visible light features are extracted with Transformer blocks alone and then fused. The experimental results can be observed that there is an obvious imbalance in the brightness adjustment between the headlights and the zebra crossing in the fused image. On the contrary, our fusion results in a high-contrast scene with good visual perception. After removing the MRF module, we simply splice the IR and visible features extracted by the encoder on the channel and feed them into the reconstruction network. The experimental results show that the whole fused scene is relatively smooth, with less gradient variation and a less vivid texture structure. Meanwhile, by observing the vehicles in the fused image, the stains on the windows, and the pedestrians on the road, it can be found that the significant target information in the infrared image is not well complemented into the fused image. There is even a situation in which the information in the IR image is lost. This situation further highlights the importance of the MEF module in integrating salient target information and background texture. This experiment shows that the designed BAU module as well as the MEF module achieved the equalization of the scene color distribution. It successfully maintains the texture details of the background area as well as the saliency of the salient targets. The ability to maintain image quality under low-light conditions is demonstrated, especially the effectiveness in retaining critical visual information.

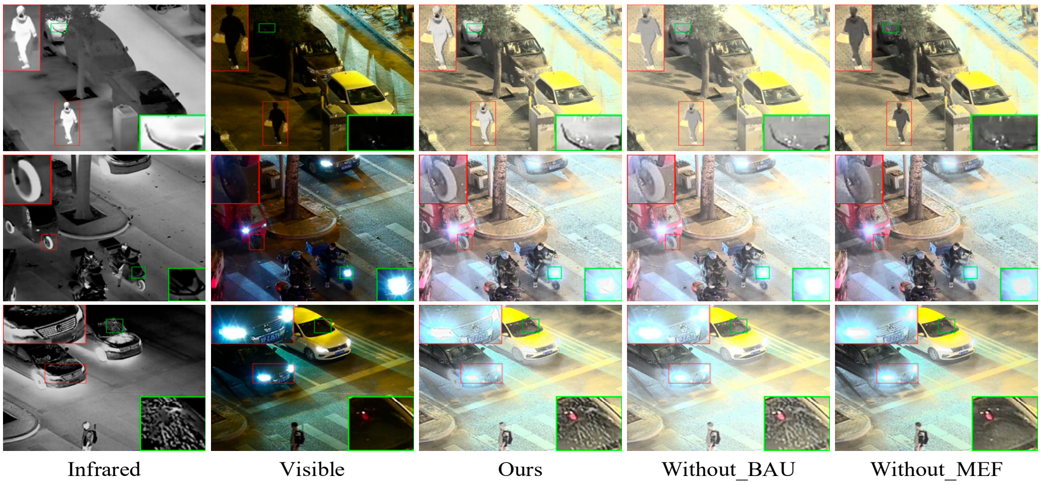


Figure 9. Visualized results of ablation on three typical infrared and visible image pairs. From top to bottom: infrared images, visible images, fused results of BMFusion, BMFusion without BAU, and BMFusion without MEF.

4.7.2. Fusion Loss Function Ablation Analysis

In order to investigate the specific effects of VGG loss and gradient loss on the quality of fused images in the fusion stage, we perform ablation experiments on VGG loss and gradient loss, respectively. It can further understand the role of each loss function in image fusion. VGG loss focuses on the overall consistency and coherence of the scene content. Gradient loss aims to optimize the network to ensure that the texture information of the source image is preserved in the fused image. As shown in Figure 10, removing VGG loss weakens the global information representation of the scene. The lack of gradient loss leads to a lack of detail representation in the fused image.

4.7.3. Overall Analysis of Ablation Experiments

In the ablation study we conducted when the BAU module is missing, there is a significant imbalance in the brightness adjustment of the fused image. This leads to an imbalance of important details and contrast in the scene. The overall visual effect of the image is affected. This imbalance is especially prominent in areas with complex lighting changes or extremely high contrast. As a result, the image appears visually too dark or too bright, lacking the desired visual level and depth. On the other hand, the fusion process is much less effective without the MEF module. This is manifested in the inadequate fusion of information between the IR and visible images. It results in an ineffective combination of critical hotspot information in the IR image and detailed texture information in the visible image. The result is that the ability of the fused image to highlight important targets and maintain environmental details is greatly reduced. This inadequate fusion will directly affect the accuracy and reliability of the fused image. Therefore, the integration of the two modules, BAU and MEF, is essential to achieving high-quality fusion of IR and visible images. This ensures excellent performance of the fused images in terms of brightness balance, information integrity, and visual effect. When gradient loss is removed, we observe a color imbalance across the scene. This resulted in the fused image tending to resemble the enhanced infrared image more. The details and texture information of the visible light image are lacking. This phenomenon indicates that gradient loss plays a crucial role in maintaining the color balance and texture details of the image. On the other hand, when VGG loss is removed, the overall sharpness of the image decreases. This is due to the fact that VGG loss is essential for maintaining the structural integrity and sharpness of the image. It ensures that the image does not lose important scene content during the fusion process by enhancing the key features and structural information of the image. Each of the two losses plays an integral role in maintaining color balance, enhancing texture details and maintaining image sharpness. In addition, we also conducted a qualitative comparison of ablation, and the specific results are shown in Table 4.

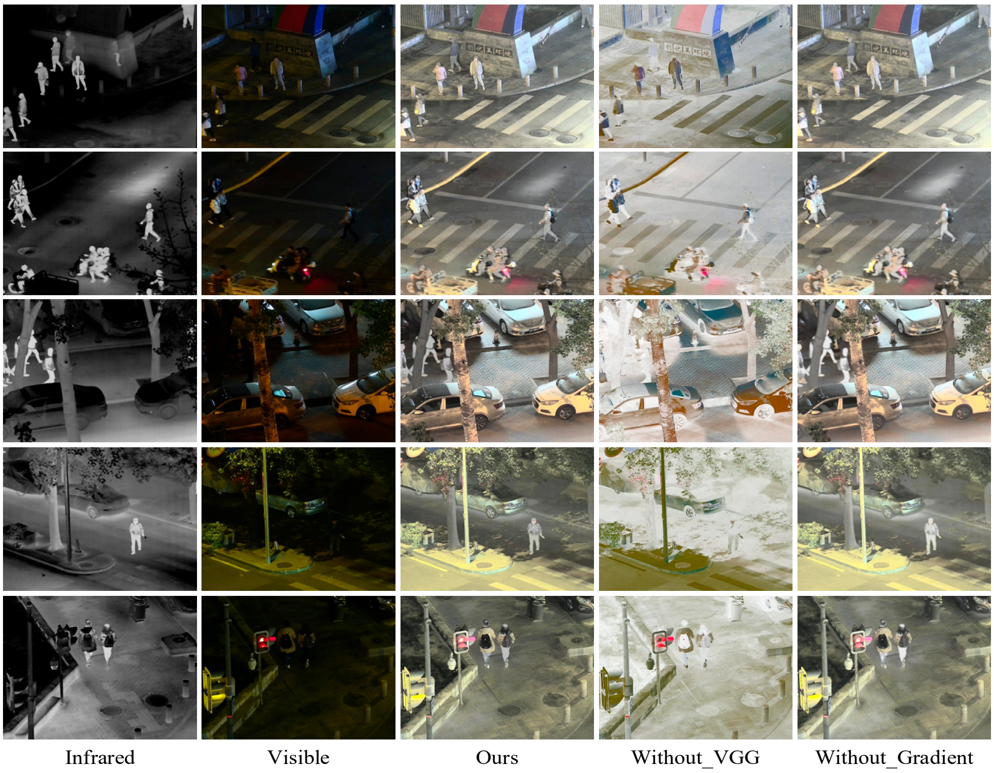


Figure 10. Visual ablation results of five typical infrared and visible light images. From top to bottom, they are: infrared image, visible light image, fusion result of BMFusion, BMFusion without VGG loss, and BMFusion without gradient loss.

Table 4. Quantitative comparison of ablation experiments.

	SD	MI	VIF	AG	EN	SF
Without_BAU	9.8232	3.6832	1.3112	6.0511	7.2856	0.0805
Without_MRF	9.7454	3.6644	1.3022	6.0001	7.2001	0.0798
Without_VGG	9.6856	3.6405	1.2904	5.9509	7.1508	0.0790
Without_Gradient	9.6179	3.6276	1.2854	5.9087	7.1001	0.0784
Ours	10.0742	3.7399	1.3379	6.1919	7.3673	0.0814

4.8. Applications in Target Detection

In order to further explore the contribution of BMFusion to advanced computer vision tasks, we take the fused images as input. Pedestrian detection experiments were conducted using the SOTA detection model YOLOv5n. The detection results were compared with the original infrared and visible light images and other different algorithms.

4.8.1. Qualitative Inorganic Experiment

We fed the infrared image, the visible light image, and the fusion results produced by each type of fusion method directly into the YOLOv5n detector, respectively. The results are shown in Figure 11. In the low-light environment, the visible image does not effectively highlight all pedestrians, and some targets are missed. The infrared image emphasizes the characters but lacks texture details, resulting in poor detection. This reflects the limitations of the two images in dark lighting conditions. Complementary information, such as high contrast, rich texture, and prominent targets in the image after fusing the two, helps to detect pedestrians effectively. However, except for BMFusion, the remaining nine fusion methods suffer from severe darkness, which weakens the complementary information of the source image. This results in the detector not being able to accurately detect all pedestrians. In contrast, BMFusion makes full use of the complementary information of infrared and visible images. It successfully overcomes the challenges of low-light environments. With its enhanced contrast and rich texture information, it provides brighter scene conditions with rich semantic information for pedestrian detection, significantly improving the accuracy of pedestrian detection.

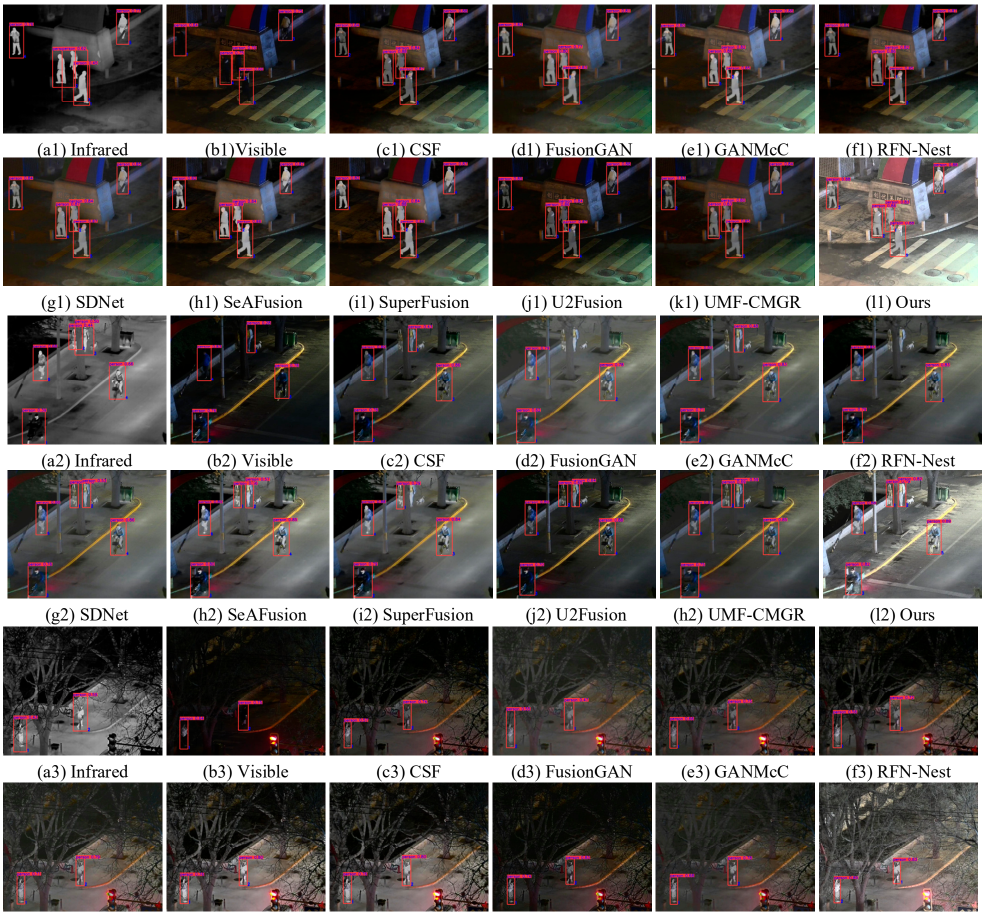

(a1) Infrared (b1)Visible (c1) CSF (d1) FusionGAN (e1) GANMcC (f1) RFN-Nest
(g1) SDNet (h1) SeAFusion (i1) SuperFusion (j1) U2Fusion (k1) UMF-CMGR (l1) Ours
(a2) Infrared (b2) Visible (c2) CSF (d2) FusionGAN (e2) GANMcC (f2) RFN-Nest
(g2) SDNet (h2) SeAFusion (i2) SuperFusion (j2) U2Fusion (h2) UMF-CMGR (l2) Ours
(a3) Infrared (b3) Visible (c3) CSF (d3) FusionGAN (e3) GANMcC (f3) RFN-Nest
(g3) SDNet (h3) SeAFusion (i3) SuperFusion (j3)U2Fusion (k3)UMF-CMGR (l3) Ours

Figure 11. Detection performance of our results with nine SOTA fusion results on different images from the LLVIP dataset.

4.8.2. Quantitative Experiment

In our study, quantitative metrics such as precision, recall, and mean accuracy (mAP) are used to evaluate the pedestrian detection task.

As shown in Table 5, the precision metric measures the proportion of positive samples correctly identified in the prediction results. Higher precision reflects fewer false detections. Recall, on the other hand, focuses on the proportion of correct detections from all actual positive samples. Higher recall means fewer missed detections. The mAP is the average detection precision under a combination of different intersection and union ratio (IoU) thresholds. It is an important indicator of the overall performance of the model. The closer the mAP value is to 1, the better the pedestrian detection is. Our fusion method achieves significant advantages in these metrics. Our fusion result, Precision's value, is at the top of the list. The value of recall is the highest. We also achieved a significant advantage in mAP. It demonstrates higher overall detection accuracy. Overall, the network designed in this paper effectively combines significant thermal target information in infrared images and texture details in visible light images. It also significantly improves the overall brightness. These properties result in higher stability of the fused images for target recognition. This results in superior performance in advanced vision tasks for target detection.

In addition, although the proposed method has shown strong potential in multimodal fusion tasks such as pedestrian re-identification, its applicability is limited due to the lack of 3D modeling, temporal information, and robustness in complex environments. Further adjustments, such as combining 3D or multi-perspective learning and improving computational efficiency [65], can significantly enhance the generalization ability and effectiveness of this method for real-world scenarios. This is also one of our future improvement goals in such tasks.

Table 5. Comparison of indicators of detection effectiveness. Bold content indicates the best indicators.

	Precision	Recall	mAP@0.50	mAP@[0.5:0.95]
VI	0.899	0.819	0.889	0.459
IR	0.725	0.631	0.709	0.287
CSF	0.928	0.872	0.937	0.564
FusionGAN	0.935	0.861	0.932	0.566
GANMcC	**0.938**	0.874	0.938	0.569
U2Fusion	0.930	0.871	0.936	0.557
RFN-Nest	0.934	0.862	0.935	0.562
SDNet	0.926	0.886	0.943	0.572
SeAFusion	0.911	0.886	0.934	0.563
SuperFusion	0.922	0.875	0.933	0.546
UMF-CMGR	0.928	0.884	0.939	0.566
Ours	0.929	**0.901**	**0.951**	**0.574**

5. Conclusions

In this paper, an innovative network architecture for the fusion of infrared and visible images under low light and complex lighting conditions is proposed. The designed fusion framework is improved in several key aspects compared to existing image fusion methods. It can be adapted to image fusion tasks in extreme lighting environments. The performance of the fused images in terms of detail retention, brightness reinforcing, and multimodal information integration achieves the desired results.

First, the BAU is designed to perform brightness adjustment and feature extraction on visible light images layer by layer. It makes it possible to effectively capture and enhance the brightness features of images in low-light environments. The BAU module adopts the combined architecture of SimAm and Transformer. It makes full use of SimAm's global brightness focusing ability and Transformer's long-range dependency modeling ability. And it ensures that the extracted features contain both local details and global consistency.

Second, this paper introduces the MRF module in the feature fusion stage. The modal features are interactively enhanced through channel-level and spatial-level attention mechanisms. It ensures that the complementary information between different modalities is preserved and enhanced in the fusion process. This multi-level feature optimization effectively enhances the expressive power of the fused features. It provides a solid foundation for the subsequent reconstruction stage.

Finally, in the feature reconstruction phase, a progressive semantically guided feature reconstruction network is used. It combines a CLIP model, a KAN module, and a neighborhood attention enhancement strategy to reconstruct the fused features layer by layer. The CLIP model provides semantic guidance during the reconstruction process for each layer. It ensures semantic consistency in the reconstruction process. The KAN module, on the other hand, refines the fusion features through adaptive channel weighting. It further enhances the complementarity and importance of different modal features. Neighbor Attention Enhancement mines the relationship between features in the reconstruction process and strengthens the detail information. It makes the fused image perform well in both global and local aspects.

Numerous experimental results show that the fusion method proposed in this paper achieves significant advantages over other state-of-the-art algorithms on publicly available datasets. Especially in low-light and nighttime environments, the performance of fused images in terms of brightness, detail retention, and visual consistency is particularly outstanding. Meanwhile, applications in advanced visual tasks also demonstrate the great potential of BM-Fusion. Overall, in this paper, through the all-round optimization of brightness enhancement, feature fusion, semantic guidance and detail reconstruction. A more stable and efficient fusion of infrared and visible images in complex environments is achieved.

Author Contributions: Formal analysis, Z.C. and C.L.; Funding acquisition, B.L.; Investigation, Z.C.; Methodology, C.L.; Resources, Z.C.; Software, C.L.; Supervision, Z.C.; Validation, C.L.; Visualization, C.L.; Writing—original draft, C.L.; Writing—review & editing, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts 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. Khan, R.; Taj, S.; Ma, X.; Noor, A.; Zhu, H.; Khan, J.; Khan, Z.U.; Khan, S.U. Advanced federated ensemble internet of learning approach for cloud based medical healthcare monitoring system. *Sci. Rep.* **2024**, *14*, 26068. [CrossRef]
2. Khan, R.; Arshad, T.; Ma, X.; Zhu, H.; Wang, C.; Khan, J.; Khan, Z.U.; Khan, S.U. GroupFormer for hyperspectral image classification through group attention. *Sci. Rep.* **2024**, *14*, 23879. [CrossRef] [PubMed]
3. Mou, J.; Gao, W.; Song, Z. Image fusion based on non-negative matrix factorization and infrared feature extraction. In Proceedings of the 2013 6th International Congress on Image and Signal Processing (CISP), Hangzhou, China, 16–18 December 2013; Volume 2, pp. 1046–1050.
4. Riley, T.; Smith, M.I. Image fusion technology for security and surveillance applications. In Proceedings of the Optics/Photonics in Security and Defence, Stockholm, Sweden, 11–16 September 2006. [CrossRef]
5. Bavirisetti, D.P.; Dhuli, R. Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Phys. Technol.* **2016**, *76*, 52–64. [CrossRef]
6. Zhang, L.; Yang, X.; Wan, Z.; Cao, D.; Lin, Y. A real-time FPGA Implementation of infrared and visible image fusion using guided filter and saliency detection. *Sensors* **2022**, *22*, 8487. [CrossRef] [PubMed]
7. Liu, Y.; Jin, J.; Wang, Q.; Shen, Y.; Dong, X. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Process.* **2014**, *97*, 9–30. [CrossRef]
8. Liu, X.; Mei, W.; Du, H. Structure tensor and nonsubsampled shearlet transform based algorithm for CT and MRI image fusion. *Neurocomputing* **2017**, *235*, 131–139. [CrossRef]
9. Choi, M.; Kim, R.Y.; Nam, M.-R.; Kim, H.O. Fusion of multispectral and panchromatic satellite images using the curvelet transform. *IEEE Geosci. Remote. Sens. Lett.* **2005**, *2*, 136–140. [CrossRef]
10. Zhang, Q.; Maldague, X. An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing. *Infrared Phys. Technol.* **2016**, *74*, 11–20. [CrossRef]
11. Wu, M.; Ma, Y.; Fan, F.; Mei, X.; Huang, J. Infrared and visible image fusion via joint convolutional sparse representation. *J. Opt. Soc. Am. A* **2020**, *37*, 1105–1115. [CrossRef] [PubMed]

12. Liu, Y.; Chen, X.; Ward, R.K.; Wang, Z.J. Image Fusion with Convolutional Sparse Representation. *IEEE Signal Process. Lett.* **2016**, *23*, 1882–1886. [CrossRef]
13. Li, H.; Wu, X.-J.; Kittler, J. MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef]
14. Cvejic, N.; Bull, D.; Canagarajah, N. Region-based multimodal image fusion using ICA bases. *IEEE Sens. J.* **2007**, *7*, 743–751. [CrossRef]
15. Fu, Z.; Wang, X.; Xu, J.; Zhou, N.; Zhao, Y. Infrared and visible images fusion based on RPCA and NSCT. *Infrared Phys. Technol.* **2016**, *77*, 114–123. [CrossRef]
16. Liu, C.; Qi, Y.; Ding, W. Infrared and visible image fusion method based on saliency detection in sparse domain. *Infrared Phys. Technol.* **2017**, *83*, 94–102. [CrossRef]
17. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17. [CrossRef]
18. Fakhari, F.; Mosavi, M.; Lajvardi, M. Image fusion based on multi-scale transform and sparse representation: An image energy approach. *IET Image Process.* **2017**, *11*, 1041–1049. [CrossRef]
19. Chen, J.; Li, X.; Luo, L.; Mei, X.; Ma, J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf. Sci.* **2019**, *508*, 64–78. [CrossRef]
20. Li, G.; Lin, Y.; Qu, X. An infrared and visible image fusion method based on multi-scale transformation and norm optimization. *Inf. Fusion* **2021**, *71*, 109–129. [CrossRef]
21. Li, X.; Wan, W.; Zhou, F.; Cheng, X.; Jie, Y.; Tan, H. Medical image fusion based on sparse representation and neighbor energy activity. *Biomed. Signal Process. Control.* **2022**, *80*, 104353. [CrossRef]
22. Li, X.; Tan, H.; Zhou, F.; Wang, G.; Li, X. Infrared and visible image fusion based on domain transform filtering and sparse representation. *Infrared Phys. Technol.* **2023**, *131*, 104701. [CrossRef]
23. Tang, D.; Xiong, Q.; Yin, H.; Zhu, Z.; Li, Y. A novel sparse representation based fusion approach for multi-focus images. *Expert Syst. Appl.* **2022**, *197*, 116737. [CrossRef]
24. Li, A.; Feng, C.; Cheng, Y.; Zhang, Y.; Yang, H. Incomplete multiview subspace clustering based on multiple kernel low-redundant representation learning. *Inf. Fusion* **2023**, *103*, 102086. [CrossRef]
25. Yu, D.; Lin, S.; Lu, X.; Wang, B.; Li, D.; Wang, Y. A multi-band image synchronous fusion method based on saliency. *Infrared Phys. Technol.* **2022**, *127*, 104466. [CrossRef]
26. Prabhakar, K.R.; Srikar, V.S.; Babu, R.V. DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4724–4732. [CrossRef]
27. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; Ma, J. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12797–12804.
28. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef] [PubMed]
29. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5009513. [CrossRef]
30. 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. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
31. Tang, W.; He, F.; Liu, Y.; Duan, Y.; Si, T. DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3159–3172. [CrossRef]
32. Tang, W.; He, F.; Liu, Y. TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation. *Pattern Recognit.* **2023**, *137*, 109295. [CrossRef]
33. Tang, L.; Zhang, H.; Xu, H.; Ma, J. 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**, *99*, 101870. [CrossRef]
34. Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; Van Gool, L. Cddfuse: Correlation-driven dualbranch feature decomposition for multi-modality image fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 5906–5916.
35. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2018**, *48*, 11–26. [CrossRef]
36. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.-P. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef] [PubMed]
37. Ma, J.; Zhang, H.; Shao, Z.; Liang, P.; Xu, H. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5005014. [CrossRef]
38. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multiscenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5792–5801.

39. Li, H.; Wu, X.-J.; Durrani, T. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [CrossRef]
40. Li, H.; Wu, X.-J.; Kittler, J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [CrossRef]
41. Jian, L.; Yang, X.; Liu, Z.; Jeon, G.; Gao, M.; Chisholm, D. SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5002215. [CrossRef]
42. Xu, H.; Wang, X.; Ma, J. DRF: Disentangled representation for visible and infrared image fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5006713. [CrossRef]
43. Xu, H.; Zhang, H.; Ma, J. Classification saliency-based rule for visible and infrared image fusion. *IEEE Trans. Comput. Imaging* **2021**, *7*, 824–836. [CrossRef]
44. Dong, X.; Wang, G.; Pang, Y.; Li, W.; Wen, J.; Meng, W.; Lu, Y. Fast efficient algorithm for enhancement of low lighting video. In Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 11–15 July 2011; pp. 1–6.
45. Guo, X.; Li, Y.; Ling, H. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **2016**, *26*, 982–993. [CrossRef]
46. Wang, S.; Zheng, J.; Hu, H.-M.; Li, B. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Trans. Image Process.* **2013**, *22*, 3538–3548. [CrossRef]
47. Ma, L.; Ma, T.; Liu, R.; Fan, X.; Luo, Z. Toward fast; flexible and robust low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5637–5646.
48. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83–84*, 79–92. [CrossRef]
49. Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; Ma, J. DIVFusion: Darkness-free infrared and visible image fusion. *Inf. Fusion* **2022**, *91*, 477–493. [CrossRef]
50. Yi, X.; Xu, H.; Zhang, H.; Tang, L.; Ma, J. Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion. *arXiv* **2024**, arXiv:2403.16387.
51. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020.
52. Knottenbelt, W.; Gao, Z.; Wray, R.; Zhang, W.Z.; Liu, J.; Crispin-Ortuzar, M. CoxKAN: Kolmogorov-Arnold Networks for Interpretable, High-Performance Survival Analysis. *arXiv* **2024**, arXiv:2409.04290.
53. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. *PMLR* **2021**, *139*, 11863–11874.
54. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. Llvip: A visible-infrared paired dataset for low-light vision. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 3489–3497.
55. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [CrossRef]
56. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518. [CrossRef]
57. Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; Ma, J. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 2121–2137. [CrossRef]
58. Di, W.; Jinyuan, L.; Xin, F.; Liu, R. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Vienna, Austria, 23–29 July 2022.
59. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 313–315. [CrossRef]
60. Cui, G.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.* **2015**, *341*, 199–209. [CrossRef]
61. Van Aardt, J.; Roberts, J.W.; Ahmed, F. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote. Sens.* **2008**, *2*, 023522–023522-28. [CrossRef]
62. Rao, Y. Recent progress in applications of in-fibre Bragg grating sensors. *Opt. Lasers Eng.* **1999**, *31*, 297–324. [CrossRef]
63. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [CrossRef]
64. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; Ren, F. Learning in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
65. Yu, Z.; Li, L.; Xie, J.; Wang, C.; Li, W.; Ning, X. Pedestrian 3D Shape Understanding for Person Re-Identification via Multi-View Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5589–5602. [CrossRef]

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.