

Article

Infrared and Visible Image Fusion for Highlighting Salient Targets in the Night Scene

Weida Zhan *, Jiale Wang, Yichun Jiang, Yu Chen , Tingyuan Zheng and Yang Hong

National Demonstration Center for Experimental Electrical, School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China

* Correspondence: zhanweida@cust.edu.cn

Abstract: The goal of infrared and visible image fusion in the night scene is to generate a fused image containing salient targets and rich textural details. However, the existing image fusion methods fail to take the unevenness of nighttime luminance into account. To address the above issue, an infrared and visible image fusion method for highlighting salient targets in the night scene is proposed. First of all, a global attention module is designed, which rescales the weights of different channels after capturing global contextual information. Second, the loss function is divided into the foreground loss and the background loss, forcing the fused image to retain rich texture details while highlighting the salient targets. Finally, a luminance estimation function is introduced to obtain the trade-off control parameters of the foreground loss function based on the nighttime luminance. It can effectively highlight salient targets by retaining the foreground information from the source images. Compared with other advanced methods, the experimental results adequately demonstrate the excellent fusion performance and generalization of the proposed method.

Keywords: image fusion; highlighting salient targets; infrared images; deep learning



Citation: Zhan, W.; Wang, J.; Jiang, Y.; Chen, Y.; Zheng, T.; Hong, Y. Infrared and Visible Image Fusion for Highlighting Salient Targets in the Night Scene. *Entropy* **2022**, *24*, 1759. <https://doi.org/10.3390/e24121759>

Academic Editor: Amelia Carolina Sparavigna

Received: 23 October 2022

Accepted: 29 November 2022

Published: 30 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Image fusion fuses multimodal images of the same scene to produce a fused image with richer information and clearer details [1,2]. Infrared and visible image fusion, as an indispensable part in image fusion direction, has been applied in all walks of life [3–5]. The fused image can be utilized as a preprocessing unit for subsequent high-level visual tasks such as object detection, object tracking and semantic segmentation. As shown in Figure 1, the salient target of a human is mainly concentrated in the thermal radiation information of the infrared image in the night scene. In addition, the uneven luminance of the night scene results in the uneven distribution of texture details in the infrared and visible images. Therefore, the fused image can effectively retain texture details while highlighting salient targets [6].



Figure 1. The infrared and visible images in the night scene. (a) Visible. (b) Infrared.

Constrained by the local receptive field of the single convolution layer, U2Fusion [7], IFCNN [8] and RXDNFuse [9] capture global contextual information by stacking plenty of convolution layers, resulting in a huge computation workload. GANMcC [10] uses the infrared and visible images as the inputs of the discriminator. It can retain the information in the source images but fails to highlight the salient targets. STDFusion [11] designs the loss function by considering only the infrared foreground and the visible background, possibly leading to the loss of some important information. Therefore, there are no comprehensive loss functions to supervise the training of the network.

To solve these problems, this paper proposes an infrared and visible image fusion method for highlighting salient targets in the night scene. The major contributions of this paper are as follows:

1. To reduce the computational burden caused by stacking convolution layers, a global attention module is added to the image fusion network. This module rescales the weights of different channels after capturing global contextual information. Assigning higher weight for the important channels makes the salient targets more prominent;
2. In order to improve the comprehensiveness of the loss function, the loss function of the proposed method is divided into the foreground loss and the background loss, forcing the fused image to contain more prominent salient targets and richer texture details;
3. To address the problem of the uneven luminance in the night scene, a luminance estimation function is introduced into the training process. It outputs the trade-off control parameters of the foreground loss function based on the nighttime luminance, which helps to retain the foreground information from source images more reasonably.

The remainder of this paper is arranged as follows. In Section 2, the reviews about infrared and visible image fusion methods and attention mechanism are presented. In Section 3, the loss function and structure of the proposed network are elaborated. The effectiveness and generalization of the proposed method are demonstrated in Section 4. Finally, the conclusion is expressed in Section 5.

2. Related Work

In this section, a literature review of infrared and visible image fusion methods and attention mechanisms are presented.

2.1. Infrared and Visible Image Fusion

Traditional image fusion methods usually use a fixed mathematical transformation to extract image features, possibly leading to a weak feature representation. Moreover, these methods do not take modal differences between multimodal images into account [12]. Deep learning-based image fusion methods, with a powerful representation capability, transform the fusion process into training different depths of neural networks.

Over the last few years, convolutional neural network-based image fusion methods have gained notable success. Xu et al. [13] adopted a dense connection structure to extract image features and added multiple jump connections between the dense blocks to increase the information flow. Tang et al. [14] proposed an image fusion network based on illumination perception, which could adaptively maintain the intensity distribution of salient targets according to light distribution.

Due to the capability of unsupervised distribution estimation, a GAN-based network is fit for image fusion tasks. Based on the former research [15], Ma et al. [16] proposed a GAN-based image fusion network with two discriminators. The infrared and visible images were adopted as the inputs of two discriminators, respectively, forcing the fused result to contain the information from both source images. Hou et al. [17] divided the input of generators into a gradient path and a contrast path and designed an adjusted image as the ground truth for the discriminator.

As the research progressed, a number of in-depth methods based on auto-encoder were proposed. Li et al. [18] adopted an encoder with nest connection structure to fully

exploit the deep features, and a hybrid attention module with channel and spatial attention was added to the fusion layer. Li et al. [19] proposed a special fusion layer with four modified residual fusion modules. The shallow and deep fusion modules were used to extract texture details and retain the significant structure, respectively. Particularly, the coding and fusion networks were trained in stages.

Although excellent fusion results can be obtained using the above methods, they still fail to effectively highlight salient targets in the fused image. The loss function of the proposed method is divided into the foreground loss and the background loss, which helps to more effectively extract and retain texture details while highlighting the salient targets.

2.2. Attention Mechanism

The attention mechanism is of great significance in the human visual system. When faced with a complex scene, the human eye tends to focus more on the important targets and ignore the irrelevant information, which helps to analyze complex scenes quickly and efficiently [20]. The attention mechanism enables selective focus on important parts by reassigning weights to input sequences [21].

The attention mechanism has achieved great success in some fields, including target detection [22], image enhancement [23], semantic segmentation [24] and emotion recognition [25]. Recently, many scholars have applied the attention mechanism to image fusion tasks. For instance, Li et al. [26] introduced a multiscale channel attention network into the generator and discriminator, respectively, to help the network focus more on salient regions in the image. Li et al. [27] argued that both channel and spatial dimensions contain rich attention information, therefore, the channel and spatial attention modules were combined for further feature extraction.

To break through the limitation of the local receptive field of the single convolution layer, Wang et al. [28] proposed a non-local attention module, which could model global contextual relationships to capture feature dependencies over long distance. However, multiple matrix multiplications were applied in this module, leading to inefficient learning of the network. In order to reduce the computational complexity, many scholars have proposed some variants based on the non-local attention module [29,30].

Despite the fact that the non-local attention module is effective in capturing global contextual information, it is unable to model the correlations between channels. Therefore, a global attention module was designed, which rescales the weights of different channels after modeling global contextual relationships.

3. The Proposed Method

In this section, the problem formulation and the core ideas are discussed. Then, the design of the loss function is presented. Finally, the structure of the image fusion network is described in detail.

3.1. Problem Formulation

A fused image can compensate for the lack of information from a single sensor [31]. The ideal result of the fused image is to retain rich texture details while highlighting the salient targets. As shown in Figure 2, in the training process, the source images were fed into the image fusion network which output the fused image. The luminance estimation function evaluated the luminance of all visible images in the training dataset and output the trade-off control parameters of the foreground loss function. Under the surveillance of the loss function, the fusion performance was improved.

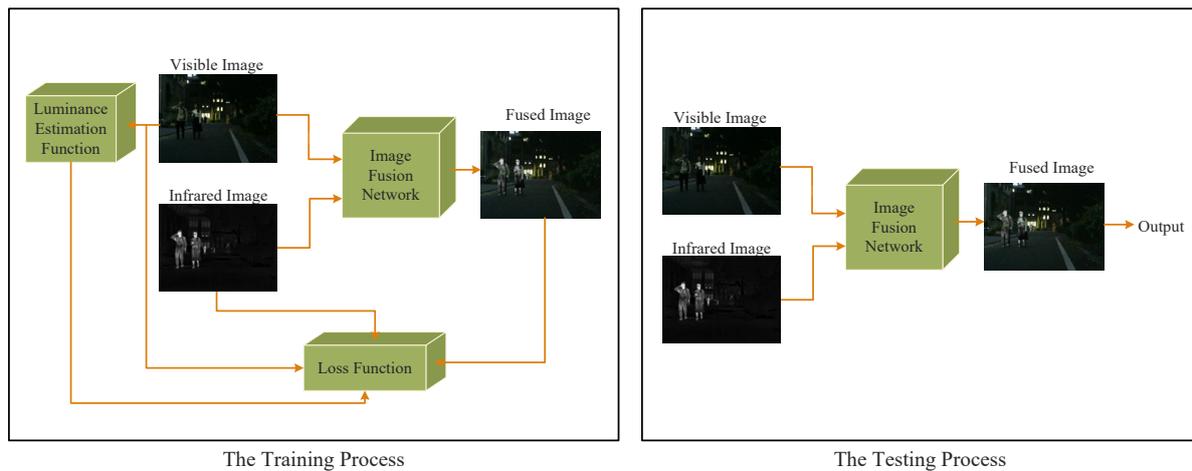


Figure 2. The entire procedure of the proposed method.

In the testing process, the final fused image was obtained by feeding the source images into the image fusion network directly without any preprocessing. The design of the loss function and the image fusion network are described in detail as follows.

3.2. Loss Function

To obtain the fused result with salient targets and rich texture details, the loss function was divided into the foreground loss and the background loss. The foreground with salient targets and the background containing rich texture details were split from the images [17].

As shown in Figure 3, humans were defined as the salient targets in this paper. First of all, the semantic segmentation results of the infrared image were acquired using Mask R-CNN [32]. The mask for humans was kept and then set as a binary image. As shown in Figure 3c,d, the foreground mask and the background mask were obtained. The source images were split into the foreground and background using the above masks [17]. It is worth mentioning that IR-Fore, VI-Fore, IR-Back and VI-Back represent the foreground and background of the infrared and visible image, respectively.

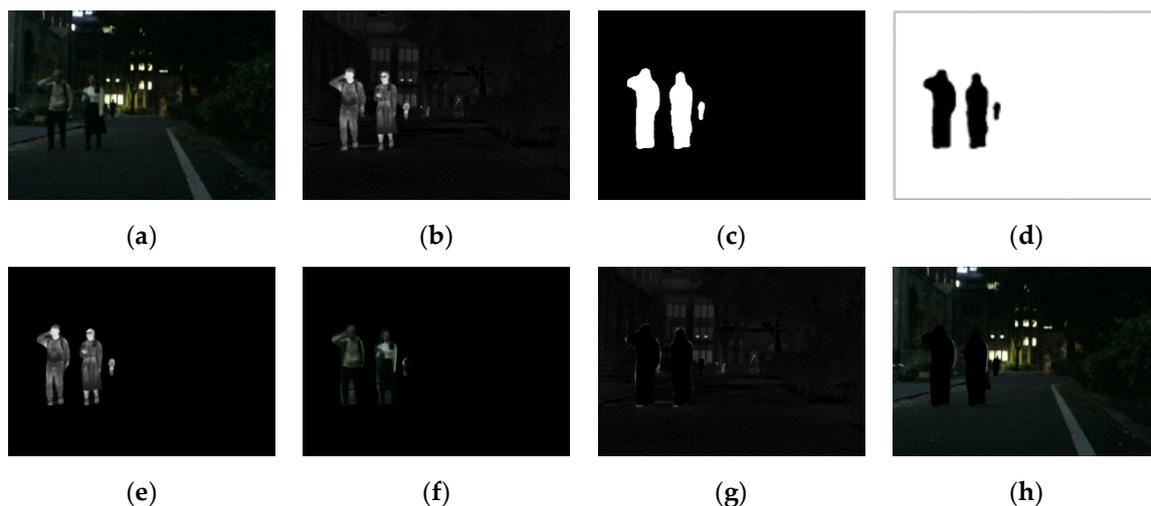


Figure 3. The source images, masks, foreground and background. (a) Visible. (b) Infrared. (c) Mask-Fore. (d) Mask-Back. (e) IR-Fore. (f) VI-Fore. (g) IR-Back. (h) VI-Back.

In Figure 3, most of the salient target information is concentrated in the infrared foreground in the night scene. However, the visible foreground also contained some salient target information. The lower the nighttime luminance, the less salient target information was contained in the visible foreground. Therefore, a luminance estimation function was

introduced to evaluate the nighttime luminance. The luminance estimation function is formulated in Equations (1) and (2).

$$\beta_2 = \frac{1}{N} \sum_{n=1}^N (1 - \log(1 + e^{-p_n})) \quad (1)$$

$$\beta_2 = 1 - \beta_1 \quad (2)$$

where β_1 and β_2 represent the trade-off control parameters of the foreground loss function. N denotes the number of visible images in the training dataset. p_n denotes the normalized average grey level value of the n th input image. According to the analysis of Equation (1), the lower the nighttime luminance, the smaller the value of β_2 .

The foreground loss function contains the foreground pixel loss $L_{pixel-fore}$ and the foreground gradient loss $L_{grad-fore}$, which are defined in Equations (3) and (4).

$$L_{pixel-fore} = \frac{\beta_1 |I_m \cdot (I_f - I_{ir})| + \beta_2 |I_m \cdot (I_f - I_{vi})|}{HW} \quad (3)$$

$$L_{grad-fore} = \frac{\beta_1 |I_m \cdot (\nabla I_f - \nabla I_{ir})| + \beta_2 |I_m \cdot (\nabla I_f - \nabla I_{vi})|}{HW} \quad (4)$$

where I_m , I_f , I_{ir} and I_{vi} represent the foreground mask, the fused image, the infrared and the visible image, respectively. H and W denote the height and width of the image. $|\cdot|$ indicates the mean absolute error. ∇ and \cdot denote the gradient operator and element multiplication operator, respectively. The fused results for different values of β_1 and β_2 are analyzed in Section 4.

The background luminance was uneven in the night scene, which lead to an uneven distribution of texture details in the source images as well. When the luminance was high, most of the texture details were concentrated in the visible background. As the luminance decreased, most of the texture details were concentrated in the infrared background. Therefore, for the background loss function, the highest pixel and gradient intensity distribution of the source images were chosen. Equations (5) and (6) for the background pixel loss $L_{pixel-back}$ and the background gradient loss $L_{grad-back}$ are shown as follows.

$$L_{pixel-back} = \frac{|(1 - I_m) \cdot (I_f - \max(I_{ir}, I_{vi}))|}{HW} \quad (5)$$

$$L_{grad-back} = \frac{|(1 - I_m) \cdot (|\nabla I_f| - \max(|\nabla I_{ir}|, |\nabla I_{vi}|))|}{HW} \quad (6)$$

where $(1 - I_{mask})$ represents the background mask and $\max(\cdot)$ denotes the element of the max selection. The total loss function L is defined in Equation (7).

$$L = \alpha_1 (L_{pixel-fore} + L_{grad-fore}) + \alpha_2 (L_{pixel-back} + L_{grad-back}) \quad (7)$$

where $\alpha_1 + \alpha_2 = 10$. The fused results for different values of α_1 and α_2 are analyzed specifically in Section 4.

3.3. Network Structure

As shown in Figure 4, the image fusion network consisted of three parts, including the feature extraction module, global attention module and image reconstruction module. First, the feature extraction module extracted the features of the source images. Then, the global attention module aggregated the global contextual information and assigned higher weight to the important channels. Finally, the image reconstruction module output the final result by reconstructing all the feature maps concatenated in the channel dimension.

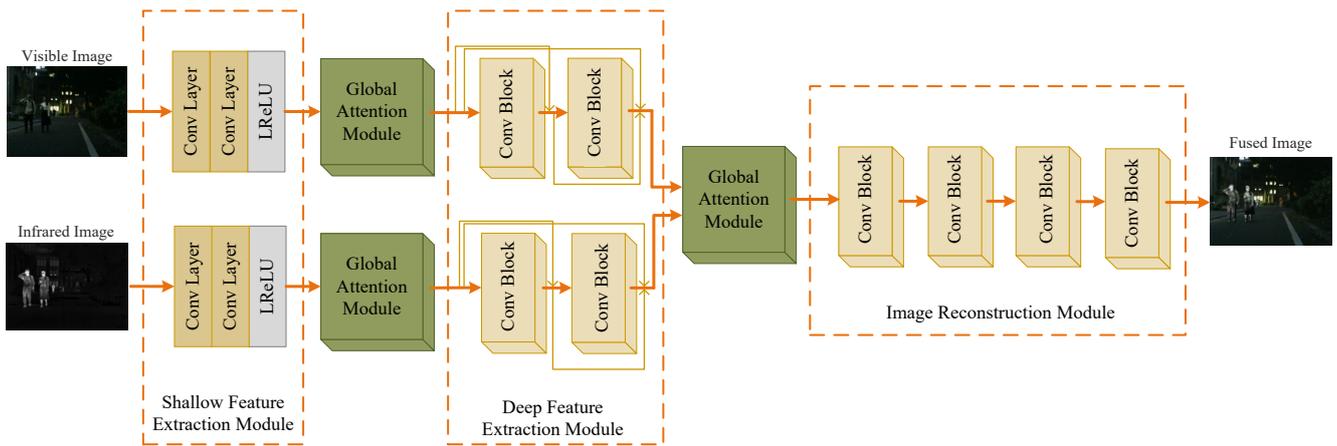


Figure 4. The structure of the image fusion network.

3.3.1. Feature Extraction Module

Since different types of source images had modal differences, the input was split into two paths. The feature extraction module for each path was designed with the same structure and respective parameters.

The feature extraction module was composed of a shallow and deep feature extraction module. In particular, the shallow feature extraction module contained two convolution layers and an activation function layer. There were two convolution blocks in the deep feature extraction module which further extracted the image features adequately. Furthermore, each block contained a convolution layer and an activation function layer. A dense connection structure was applied between each block to ensure the full utilization of information. It is necessary to note that the padding of each convolution layer was set to *SAME* and the stride was set to one. The parameter settings of the feature extraction module are shown in Table 1.

Table 1. Parameter settings of the feature extraction module.

Module	Layer	Input Channel	Output Channel	Activation
Shallow Feature Extraction Module	1	1	16	None
	2	16	16	Leaky ReLU
Deep Feature Extraction Module	1	16	32	Leaky ReLU
	2	32	64	Leaky ReLU

3.3.2. Global Attention Module

Two global attention modules were introduced after the shallow and deep feature extraction modules. First, the global context was modeled to capture long-range feature dependencies. Later, the correlation between each channel was built through a fully connected layer. The structure of the global attention module is shown in Figure 5.

For the input feature $\Phi^I \in R^{C \times H \times W}$, it was first reshaped and transposed to $\Phi^X \in R^{C \times HW}$. Then, Φ^I was reshaped to $\Phi^Y \in R^{HW \times 1 \times 1}$ after a convolution layer. A softmax operation was applied to calculate the attention map $A \in R^{C \times 1 \times 1}$, which is defined by:

$$A = \frac{\exp(\Phi^Y)}{\sum_{i=1}^{HW} \exp(\Phi^Y)} \tag{8}$$

where $i \in [1, HW]$. Subsequently, a matrix multiplication operation between A and Φ^X

was performed, obtaining the global attention map $\Phi^G \in R^{C \times 1 \times 1}$, which is defined by:

$$\Phi^G = A \times \Phi^X \tag{9}$$

where \times denotes the matrix multiplication operation. After obtaining the global context dependencies, the channel relationship was modeled with a fully connected layer. Then, a softmax operation was applied to obtain the global attention map $G \in R^{C \times 1 \times 1}$, which is defined by:

$$G = \sigma(f_c(W)) \tag{10}$$

where $f_c(\cdot)$ denotes the fully connected layer, W is the parameter of the fully connected layer and $\sigma(\cdot)$ denotes the Sigmoid operation. Finally, the final output feature map $\Phi^O \in R^{C \times H \times W}$ was obtained by applying the corresponding pixel multiplication operation between G and Φ^I . The final output feature map Φ^O is formulated as follows.

$$\Phi^O = G \times \Phi^I \tag{11}$$

where \times denotes a corresponding pixel multiplication operation.

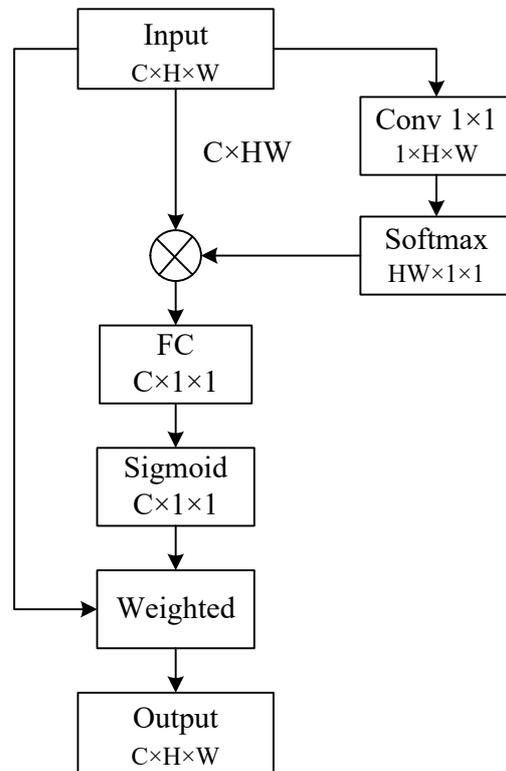


Figure 5. Structure of the global attention module.

3.3.3. Image Reconstruction Module

The image reconstruction module reconstructed the whole feature maps concatenated in the channel dimension and output the final result. There were four convolution blocks in this module. In particular, each block contained a convolution layer and an activation function layer. Moreover, the padding of each convolution layer was set to *SAME* and the stride was set to one. The image fusion network did not change the size of the input. The parameter settings of the image reconstruction module are shown in Table 2.

Table 2. Parameter settings of the image reconstruction module.

Module	Layer	Input Channel	Output Channel	Activation
Image Reconstruction Module	1	224	128	Leaky ReLU
	2	128	64	Leaky ReLU
	3	64	32	Leaky ReLU
	4	32	1	Leaky ReLU

4. Experiments

The effectiveness and generalization of the proposed method are demonstrated in this section. First of all, the experimental details are presented. Then, the experimental results of two datasets are discussed in detail. Finally, ablation experiments concerning the global attention module, the feature extraction module and the trade-off control parameter are presented.

4.1. Experimental Details

4.1.1. Training Dataset

The MSRS dataset [14] was chosen to train the network model. This dataset was a mixture of daytime and nighttime road-related scenes, containing various elements such as pedestrians, vehicles, buildings, etc. One hundred and eighty-eight pairs of the nighttime images were selected from the dataset. The foreground mask and the background mask were obtained through the modified semantic segmentation network Mask R-CNN [32]. Each selected image was cropped into image patches with the size of 120×120 by setting the stride to 64, resulting in 13,160 patch pairs for the network training. In the testing process, 26 and 16 pairs of images were chosen from the MSRS dataset and the M³FD dataset, respectively, for comparison experiments. Note that during the testing process, the input did not require any preprocessing.

4.1.2. Training Details

The network model was implemented on the TensorFlow platform, and Adam was used as the optimization solver. The batch size and epoch were set to 32 and 30, respectively. In addition, the learning rate was set to 10×10^{-3} . Particularly, we set $\alpha_1 = 4$, $\alpha_2 = 6$, $\beta_1 = 0.66$ and $\beta_2 = 0.34$.

This paper evaluates the quality of fused results by two means containing qualitative and quantitative comparisons. Six compared methods with the code and default parameters were chosen, including GTF [33], DenseFuse [34], U2Fusion [7], GANMcC [10], RFN-Nest [19] and STDFusion [11]. Six typical evaluation metrics were selected including entropy (EN) [35], mutual information (MI) [36], standard deviation (SD) [37], spatial frequency (SF) [38], visual information fidelity (VIF) [39] and quality of images ($Q^{ab/f}$) [40]. It is worth mentioning that the larger the above evaluation metrics, the better the fused result.

4.2. Results of the MSRS Dataset

For assessing the fusion performance of the proposed method in a comprehensive manner, the MSRS dataset was chosen for comparison with six state-of-the-art methods.

4.2.1. Qualitative Results

For visually comparing the fusion performance of different methods, four pairs of images from the MSRS dataset were chosen for qualitative comparison. Notably, in Figures 6–9, red and green boxes were used to mark the salient target and texture details, respectively. All methods achieved excellent performance in retaining both salient targets and texture details. Specifically, as shown in Figure 6, DenseFuse [34], U2Fusion [7], GANMcC [10] and STDFusion [11] were able to highlight the salient targets well, but the texture details of the building's windows in the background were not clear. GTF [33] and

RFN-Nest [19] retained the thermal radiation information well, but the edges of the salient targets were blurred.

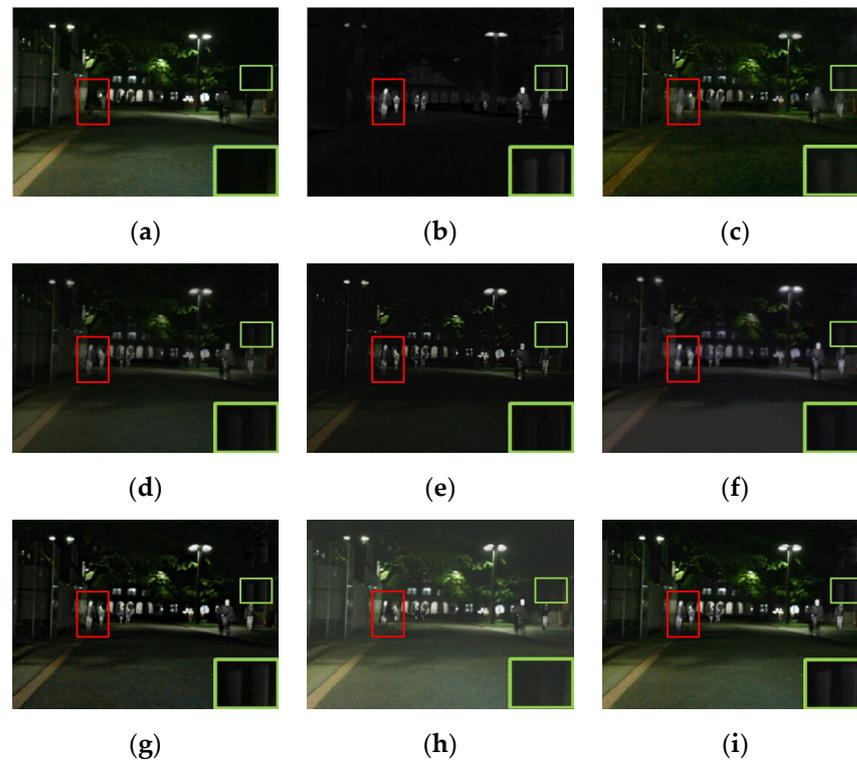


Figure 6. The qualitative results of the first scene of the MSRS dataset. (a) Visible. (b) Infrared. (c) GTF. (d) DenseFuse. (e) U2Fusion. (f) GANMcC. (g) RFN-Nest. (h) STDFusion. (i) Ours.

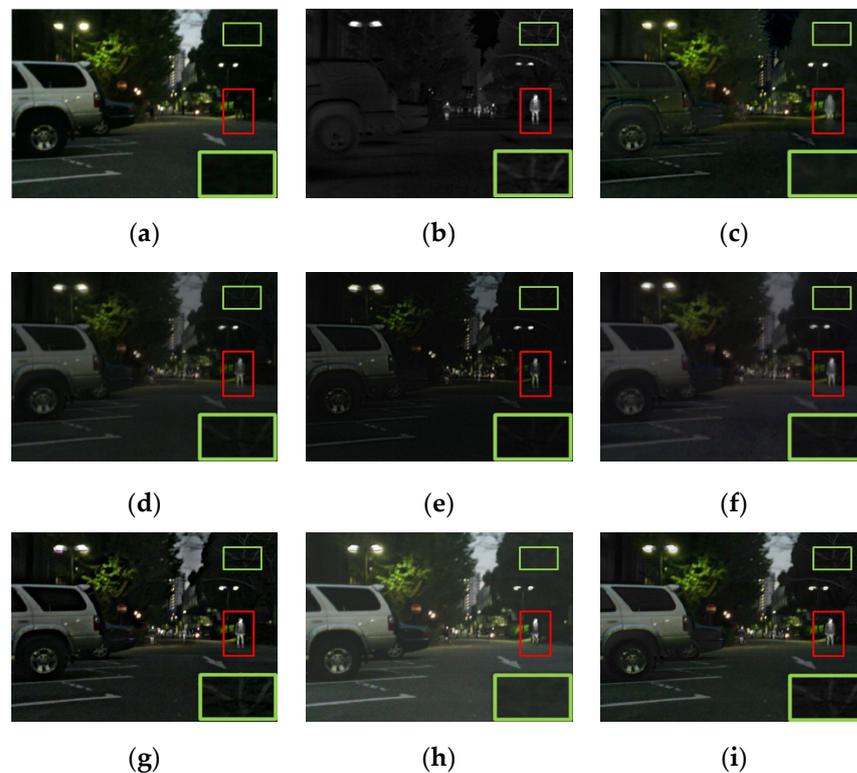


Figure 7. The qualitative results of the second scene of the MSRS dataset. (a) Visible. (b) Infrared. (c) GTF. (d) DenseFuse. (e) U2Fusion. (f) GANMcC. (g) RFN-Nest. (h) STDFusion. (i) Ours.

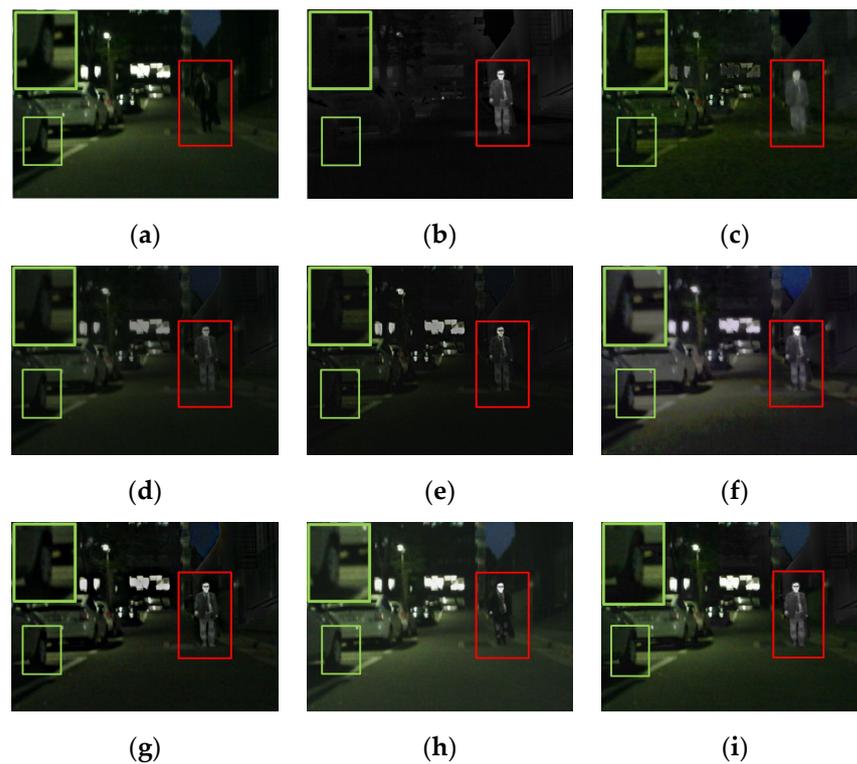


Figure 8. The qualitative results of the third scene of the MSRS dataset. (a) Visible. (b) Infrared. (c) GTF. (d) DenseFuse. (e) U2Fusion. (f) GANMcC. (g) RFN-Nest. (h) STDFusion. (i) Ours.

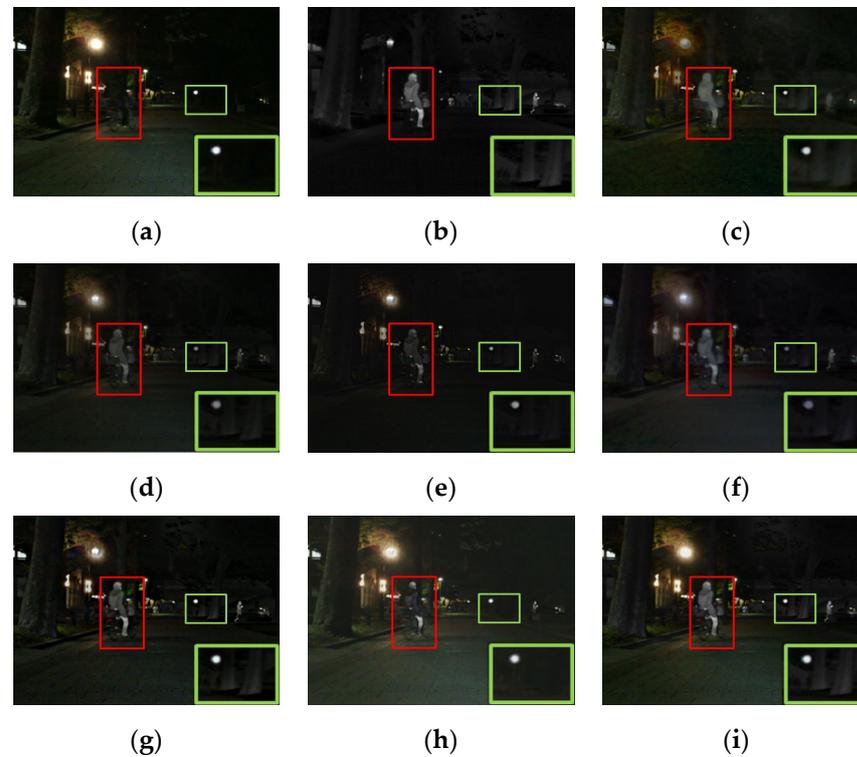


Figure 9. The qualitative results of the fourth scene of the MSRS dataset. (a) Visible. (b) Infrared. (c) GTF. (d) DenseFuse. (e) U2Fusion. (f) GANMcC. (g) RFN-Nest. (h) STDFusion. (i) Ours.

For the other three scenes, the proposed method effectively highlighted the salient targets. In addition, it also had some advantages in retaining texture details in the background. To be specific, our fused result had the sharpest detail of the tree branches in Figure 7. As

shown in Figure 8, the outline of the car wheels was clear. Moreover, the tree trunk could be clearly distinguished from its surroundings in Figure 9. The above results indicated that the proposed method had better capabilities in highlighting salient targets and preserving rich texture details.

4.2.2. Quantitative Results

For further evaluating the effectiveness of different methods, 26 pairs of images from the MSRS dataset were chosen for quantitative comparison. As shown in Figure 10 and Table 3, the proposed method had a remarkable advantage in five metrics, including EN, MI, SD, SF and $Q^{ab/f}$. The value of the VIF for the proposed method was second to the value for RFN-Nest by a narrow margin. Specifically, the best results for EN and MI showed that the fused results contained a whole wealth of information transferred from the source images. The maximum values of the SD and SF suggested that our fused image had a relatively high contrast ratio. Even if the value of the VIF was not the best, the fused result still effectively highlighted the salient targets and preserved rich texture details.

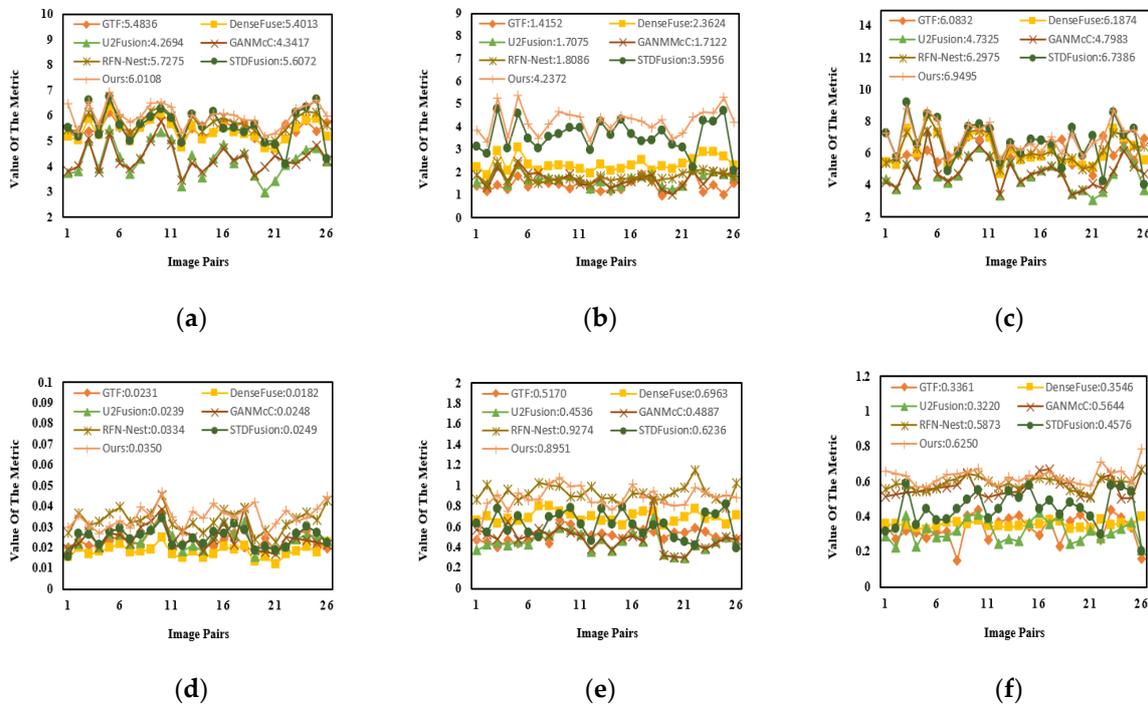


Figure 10. The quantitative results of the MSRS dataset. (a) EN. (b) MI. (c) SD. (d) SF. (e) VIF. (f) $Q^{ab/f}$.

Table 3. The quantitative results of the MSRS dataset. Red and blue indicate the best and the second-best results, respectively.

Metric	GTF	DenseFuse	U2Fusion	GANMcC	RFN-Nest	STDFusion	Ours
EN	5.4836	5.4013	4.2694	4.3417	5.7275	5.6072	6.0108
MI	1.4152	2.3624	1.7075	1.7122	1.8086	3.5956	4.2372
SD	6.0832	6.1874	4.7325	4.7983	6.2975	6.7386	6.9495
SF	0.0231	0.0182	0.0239	0.0248	0.0334	0.0249	0.0350
VIF	0.5170	0.6963	0.4536	0.4887	0.9274	0.6236	0.8951
$Q^{ab/f}$	0.3361	0.3546	0.3220	0.5644	0.5874	0.4576	0.6250

4.3. Results of the M³FD Dataset

For further evaluating the generalization of the proposed method, the M³FD dataset was used to test the network model trained using the MSRS dataset.

4.3.1. Qualitative Results

A pair of images was chosen from the M³FD dataset for qualitative comparison. As shown in Figure 11, all methods retained the infrared thermal radiation information as well as the texture details. Specifically, the fused images of the proposed method were rich in thermal radiation information. Both near and distant targets could be easily distinguished from the background. Compared with other methods, the textural details of the proposed method were much clearer and more suitable for visual observation, for instance the detail of the building in Figure 11.



Figure 11. The qualitative results of the M³FD dataset. (a) Visible. (b) Infrared. (c) GTF. (d) DenseFuse. (e) U2Fusion. (f) GANMcC. (g) RFN-Nest. (h) STDFusion. (i) Ours.

The excellent results demonstrate the generalization of the proposed method. To sum up, the fused results were effective in highlighting salient targets when applied in the night scene, and the texture details were clearly visible.

4.3.2. Quantitative Results

Sixteen pairs of images from the M³FD dataset were chosen for quantitative comparison. As shown in Figure 12 and Table 4, the proposed method obtained the maximum mean values for four evaluation metrics including EN, MI, VIF and $Q^{ab/f}$. This indicated that our fused images were rich in information and were more suitable for observation with the human visual system. Although our proposed method did not achieve the best results for the SD and SF, the fused images still had a sufficiently high contrast. The above experimental results adequately prove the excellent fusion performance and outstanding generalization.

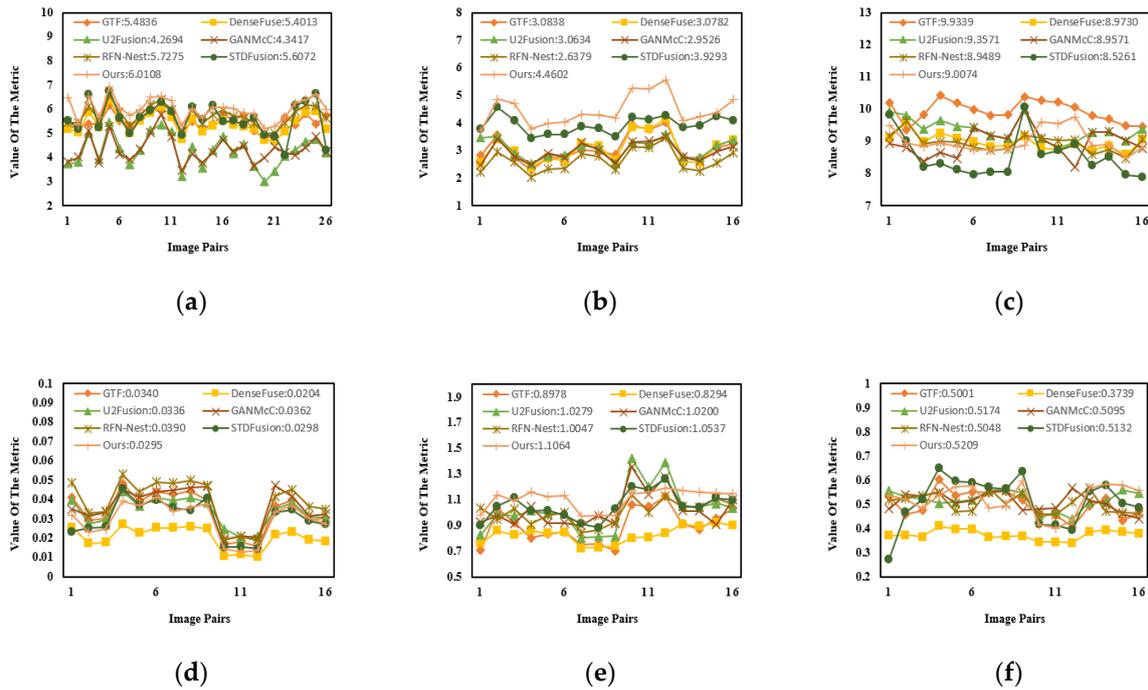


Figure 12. The quantitative results of the M³FD dataset. (a) EN. (b) MI. (c) SD. (d) SF. (e) VIF. (f) Q^{ab/f}.

Table 4. The quantitative results of the M³FD dataset. Red and blue indicate the best and the second-best results, respectively.

Metric	GTF	DenseFuse	U2Fusion	GANMcC	RFN-Nest	STDFusion	Ours
EN	6.5086	6.6316	7.0221	6.9833	6.7907	6.7776	7.1993
MI	3.0838	3.0782	3.0634	2.9526	2.6379	3.9293	4.4602
SD	9.9339	8.9730	9.3571	8.9571	8.9489	8.5261	9.0074
SF	0.0340	0.0204	0.0336	0.0362	0.0390	0.0298	0.0295
VIF	0.8978	0.8294	1.0279	1.0200	1.0047	1.0537	1.1064
Q ^{ab/f}	0.5001	0.3739	0.5174	0.5095	0.5048	0.5132	0.5209

4.4. Ablation Experiments

4.4.1. Global Attention Module

The perceptual field of the single convolution layer is limited, so the contextual information can only be captured by stacking plenty of convolution layers. To address the huge computational effort, the global attention module was designed. This module modeled the global contextual relationships and then rescaled the weights of different channels to make salient targets more prominent. For confirming the usefulness of the global attention module, three ablation experiments were designed, involving removing the first and the second global attention module, respectively, and removing both global attention modules.

One hundred pairs of images from the M³FD dataset were chosen for qualitative and quantitative analyses. In Figure 13, for the fused result with the single global attention module, the salient targets were relatively blurred. The fused result without the global attention module performed the worst in highlighting salient targets. In Table 5, the maximum values for five evaluation metrics were achieved when two global attention modules were retained. Therefore, the global attention module assigned higher weights to the significant channels, resulting in the effective highlighting of the salient targets.

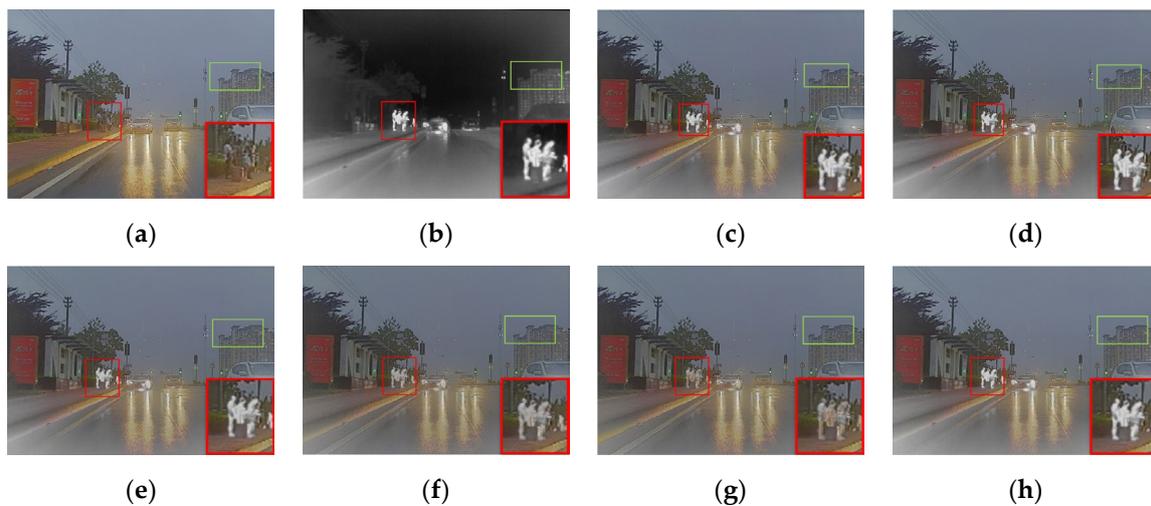


Figure 13. The qualitative results of ablation experiments concerning the global attention module. (a) Visible. (b) Infrared. (c) Replacing attention module with SENet. (d) Replacing attention module with ECANet. (e) Without the 1st attention module. (f) Without the 2nd attention module. (g) Without both attention modules. (h) Ours.

Table 5. The quantitative results of the effectiveness of the global attention module. Red and blue indicate the best and the second-best results, respectively.

Metric	Without the 1st Attention Module	Without the 2nd Attention Module	Without Both Attention Modules	Ours
EN	6.7450	6.7584	6.6899	6.7844
MI	4.4664	4.5162	4.4048	4.7104
SD	8.9386	8.9254	8.8987	8.9747
SF	0.0320	0.0271	0.0220	0.0294
VIF	0.9936	0.9921	0.9836	1.0026
$Q^{ab/f}$	0.5010	0.4927	0.4793	0.5199

In order to further verify its advancement, the global attention module was replaced with SENet [41] and ECANet [42]. As shown in Figure 13, all global attention modules effectively highlighted the salient targets. Therefore, it was difficult to judge the superiority of these modules with human visual inspection. Therefore, 100 pairs of images from the M³FD dataset were utilized for further quantitative comparison. In Table 6, all evaluation metrics of the proposed module achieved maximum values. Therefore, all results confirmed its advancement.

Table 6. The quantitative results of the advancement of the global attention module. Red and blue indicate the best and the second-best results, respectively.

Metric	Replacing Attention Module with SENet	Replacing Attention Module with ECANet	Ours
EN	6.5809	6.6445	6.7844
MI	4.6827	4.6111	4.7104
SD	8.9298	8.9390	8.9747
SF	0.0277	0.0283	0.0294
VIF	0.9867	0.9370	1.0026
$Q^{ab/f}$	0.5170	0.5007	0.5199

4.4.2. Feature Extraction Module

The global attention module effectively avoided the problem of high computational complexity caused by stacking convolution layers. To further verify the effect of the number of convolution blocks in the deep feature extraction module, two ablation experiments were

designed. Three or four convolution blocks were adopted in the deep feature extraction module. The output channels of the third and fourth convolution layer were set up to 96 and 128, respectively. The other parameters were set up as before. One hundred pairs of images from the M³FD dataset were chosen for quantitative analysis. As shown in Table 7, the proposed method achieved the maximum values for four evaluation metrics including EN, MI, SD and VIF. Note that the evaluation metrics did not rise with the increase in convolution blocks, which meant that the network was possibly overfitted. Therefore, only two convolution blocks were adopted in the deep feature extraction module.

Table 7. The quantitative results of the effect of the number of convolution blocks. Red and blue indicate the best and the second-best results, respectively.

Metric	3-Conv-Block	4-Conv-Block	Ours (2-Conv-Block)
EN	6.7299	6.3615	6.7844
MI	4.6058	3.0175	4.7104
SD	8.9157	8.7353	8.9747
SF	0.0293	0.0296	0.0294
VIF	0.9843	0.8739	1.0026
Q ^{ab/f}	0.5310	0.4540	0.5199

4.4.3. Trade-Off Control Parameters

For retaining the salient targets and rich texture details at the same time, the loss function was divided into the foreground loss and the background loss. The trade-off control parameters of the foreground loss were α_1 and α_2 , where $\alpha_1 + \alpha_2 = 10$. Fixing the other parameters, we set $\alpha_1 = 1, 2, \dots, 9$. One hundred pairs of images from the M³FD dataset were selected for analysis. The variation trends in the evaluation metrics under different values of α_1 can be observed in Figure 14. In Table 8, the best results were obtained for all evaluation metrics except for Q^{ab/f} when α_1 was set to four.

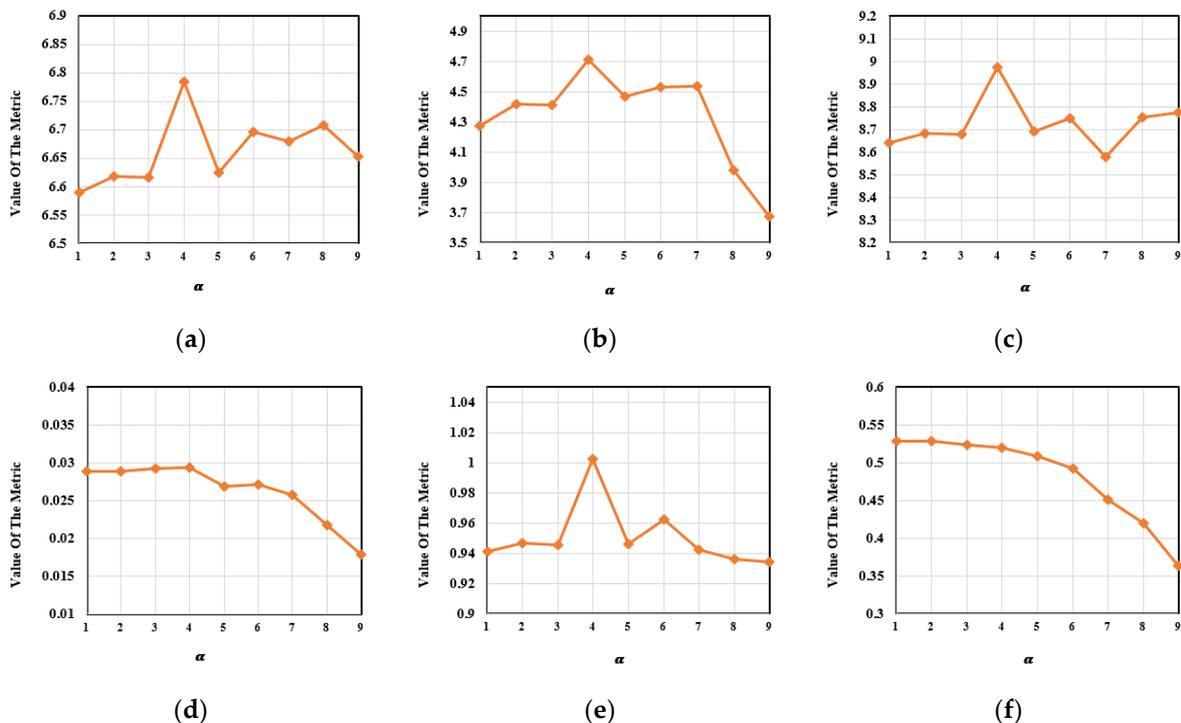


Figure 14. The variation trends in the metrics for the trade-off control parameter α . (a) EN. (b) MI. (c) SD. (d) SF. (e) VIF. (f) Q^{ab/f}.

Table 8. The ablation experimental results for the trade-off control parameter α . Red and blue indicate the best and the second-best results, respectively.

Metric	$\alpha_1 = 1$	$\alpha_1 = 2$	$\alpha_1 = 3$	$\alpha_1 = 5$	$\alpha_1 = 6$	$\alpha_1 = 7$	$\alpha_1 = 8$	$\alpha_1 = 9$	Ours ($\alpha_1 = 4$)
EN	6.5890	6.6176	6.6162	6.6245	6.6959	6.6796	6.7081	6.6526	6.7844
MI	4.2757	4.4161	4.4141	4.4687	4.5292	4.5340	3.9771	3.6744	4.7104
SD	8.6407	8.6807	8.6769	8.6910	8.7493	8.5791	8.7539	8.7759	8.9747
SF	0.0289	0.0289	0.0293	0.0269	0.0271	0.0257	0.0218	0.0179	0.0294
VIF	0.9412	0.9467	0.9455	0.9462	0.9625	0.9423	0.9358	0.9341	1.0026
$Q^{ab/f}$	0.5290	0.5287	0.5240	0.5082	0.4921	0.4509	0.4201	0.3629	0.5199

By evaluating the luminance of all visible images in the training dataset, the luminance estimation function output the trade-off control parameters β_1 and β_2 , where $\beta_1 + \beta_2 = 1$. Fixing the other parameters in the network, we set $\beta_1 = 0.1, 0.2, \dots, 0.9$. One hundred pairs of images from the M³FD dataset were selected for qualitative and quantitative analyses. In Figure 15, the variation trends in the evaluation metrics under the different values of β_1 are shown. As shown in Table 9, our proposed method acquired the best results for four evaluation metrics including EN, MI, SD and VIF.

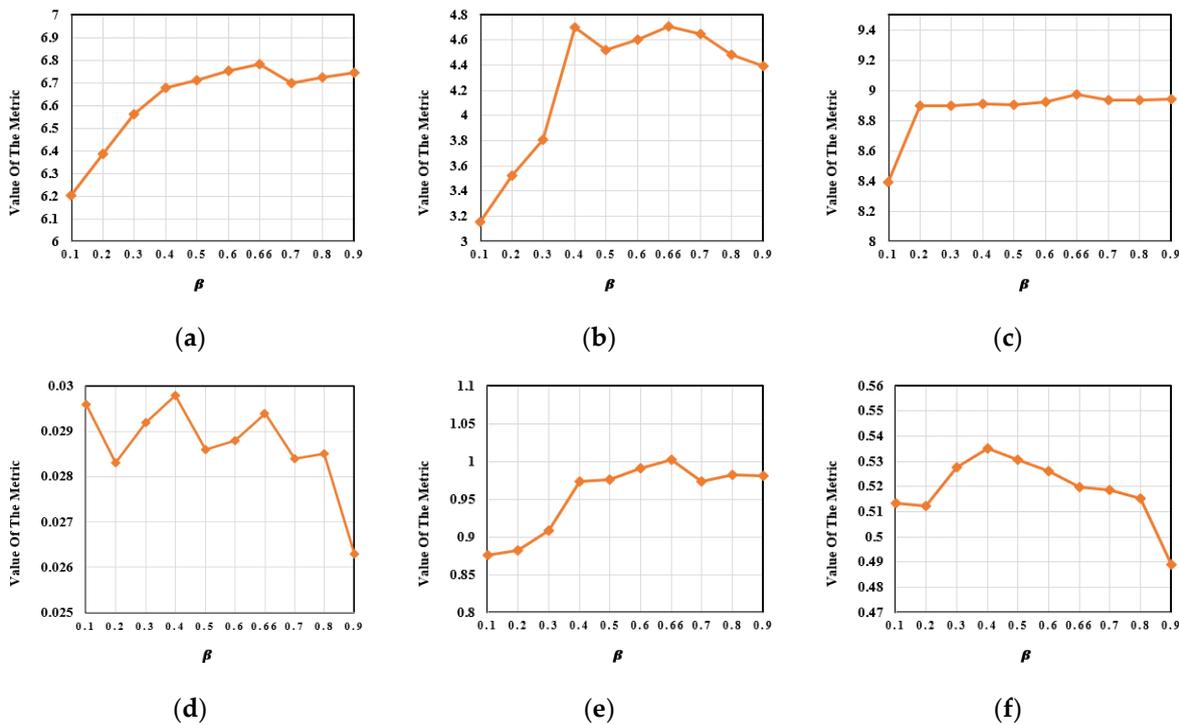


Figure 15. The variation trends in the metrics for the trade-off control parameter β . (a) EN. (b) MI. (c) SD. (d) SF. (e) VIF. (f) $Q^{ab/f}$.

In order to further verify the validity of the luminance estimation function, the experiment was continued with a reduction in the step length. We set $\beta_1 = 0.61, 0.62, \dots, 0.69$ and kept the other parameters in the network unchanged.

One hundred pairs of images from the M³FD dataset were chosen for quantitative analysis using the above 10 sets of ablation experiments. In Figure 16, the variation trends in the evaluation metrics under the different values of β_1 are shown. As shown in Table 10, our proposed method acquired the best results for four evaluation metrics including EN, SD, SF and VIF. In summary, the luminance estimation function could effectively evaluate the luminance of the night scene and output a reasonable value of β .

Table 9. The ablation experimental results for the trade-off control parameter β . Red and blue indicate the best and the second-best results, respectively.

Metric	$\beta_1 = 0.1$	$\beta_1 = 0.2$	$\beta_1 = 0.3$	$\beta_1 = 0.4$	$\beta_1 = 0.5$	$\beta_1 = 0.6$	$\beta_1 = 0.7$	$\beta_1 = 0.8$	$\beta_1 = 0.9$	Ours ($\beta_1 = 0.66$)
EN	6.2031	6.3875	6.5618	6.6762	6.7122	6.7530	6.7006	6.7238	6.7444	6.7844
MI	3.1523	3.5214	3.8086	4.7037	4.5235	4.6019	4.6483	4.4814	4.3923	4.7104
SD	8.3903	8.8980	8.8988	8.9110	8.9037	8.9236	8.9331	8.9339	8.9419	8.9747
SF	0.0296	0.0283	0.0292	0.0298	0.0286	0.0288	0.0284	0.0285	0.0263	0.0294
VIF	0.8754	0.8825	0.9089	0.9730	0.9754	0.9915	0.9735	0.9819	0.9807	1.0026
$Q^{ab/f}$	0.5133	0.5124	0.5275	0.5350	0.5308	0.5261	0.5187	0.5154	0.4890	0.5199

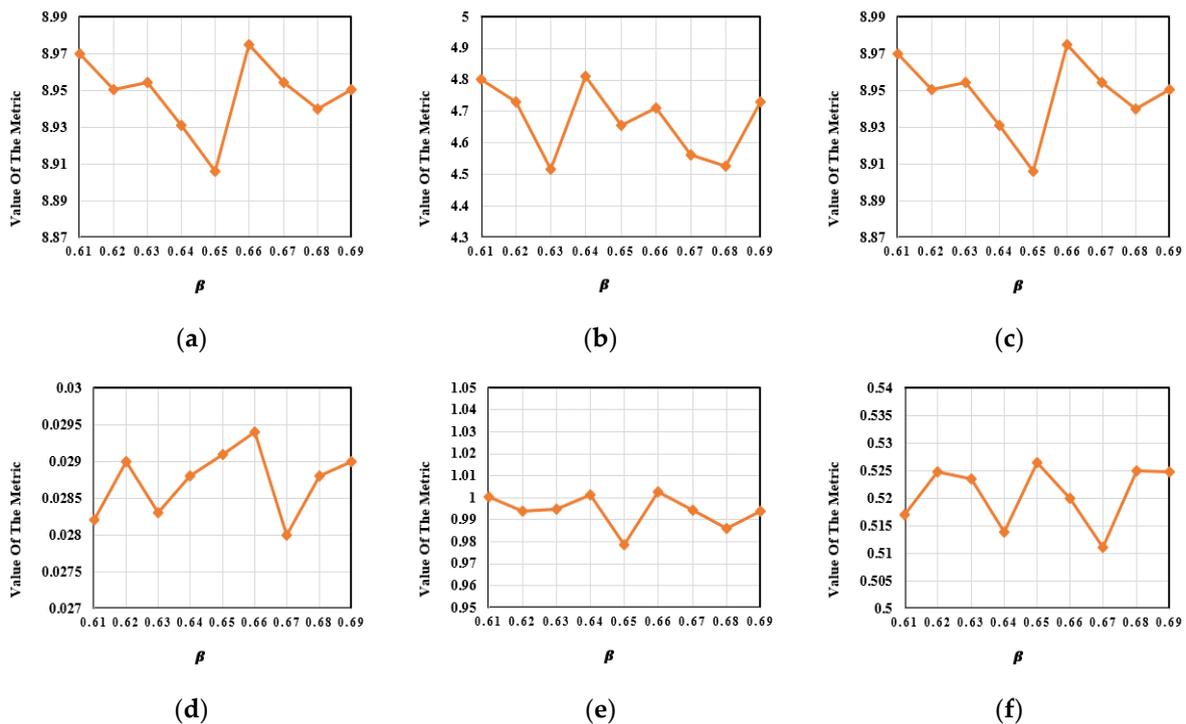


Figure 16. The further ablation experimental results for the trade-off control parameter β . (a) EN. (b) MI. (c) SD. (d) SF. (e) VIF. (f) $Q^{ab/f}$.

Table 10. The ablation experimental results for the trade-off control parameter β . Red and blue indicate the best and the second-best results, respectively.

Metric	$\beta_1 = 0.61$	$\beta_1 = 0.62$	$\beta_1 = 0.63$	$\beta_1 = 0.64$	$\beta_1 = 0.65$	$\beta_1 = 0.67$	$\beta_1 = 0.68$	$\beta_1 = 0.69$	Ours ($\beta_1 = 0.66$)
EN	6.7816	6.7521	6.7838	6.7537	6.7061	6.7635	6.7244	6.7521	6.7844
MI	4.8015	4.5154	4.8101	4.6564	4.5619	4.5265	4.5265	4.7309	4.7104
SD	8.9698	8.9504	8.9545	8.9312	8.9061	8.9546	8.9399	8.9504	8.9747
SF	0.0282	0.0290	0.0283	0.0288	0.0291	0.0280	0.0288	0.0290	0.0294
VIF	1.0003	0.9940	0.9946	1.0012	0.9788	0.9944	0.9861	0.9940	1.0026
$Q^{ab/f}$	0.5170	0.5248	0.5235	0.5138	0.5264	0.5111	0.5250	0.5248	0.5199

As shown in Figure 17, the salient targets were very prominent and the details of the building were very distinct when α_1 and β_1 were set to 4 and 0.66, respectively. All the above experiments adequately prove the rationality of the parameter settings in this paper.

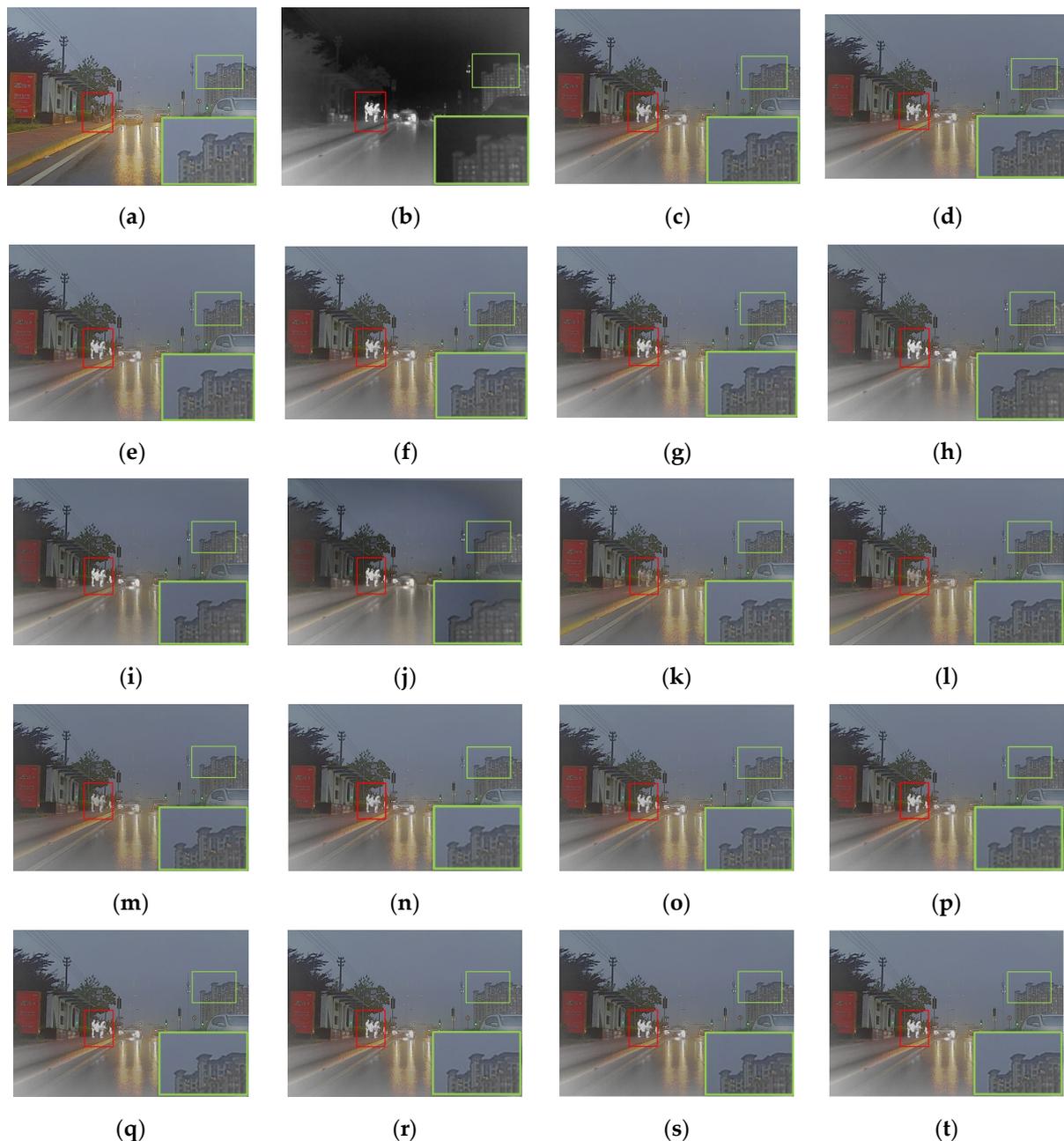


Figure 17. The ablation experimental results for the trade-off control parameters α and β . (a) Visible. (b) Infrared. (c) $\alpha_1 = 1$. (d) $\alpha_1 = 2$. (e) $\alpha_1 = 3$. (f) $\alpha_1 = 5$. (g) $\alpha_1 = 6$. (h) $\alpha_1 = 7$. (i) $\alpha_1 = 8$. (j) $\alpha_1 = 9$. (k) $\beta_1 = 0.1$. (l) $\beta_1 = 0.2$. (m) $\beta_1 = 0.3$. (n) $\beta_1 = 0.4$. (o) $\beta_1 = 0.5$. (p) $\beta_1 = 0.6$. (q) $\beta_1 = 0.7$. (r) $\beta_1 = 0.8$. (s) $\beta_1 = 0.9$. (t) Ours ($\alpha_1 = 4$, $\beta_1 = 0.66$).

5. Conclusions

This paper proposed an infrared and visible image fusion method for highlighting salient targets in the night scene. First, a global attention module was designed, which rescaled the weights of different channels after capturing the global contextual information. Second, the loss function was divided into the foreground loss and the background loss, forcing the fused image to retain rich texture details while highlighting the salient targets. Moreover, a luminance estimation function was introduced to obtain the trade-off control parameters of the foreground loss function based on the nighttime luminance. It effectively highlighted salient targets by retaining foreground information from the source images. In

summary, the effectiveness and generalization of the proposed method was demonstrated by comparing with other advanced methods.

Author Contributions: Data curation, Y.C. and T.Z.; formal analysis, J.W. and Y.J.; funding acquisition, W.Z.; investigation, Y.H.; methodology, J.W.; project administration, W.Z.; software, J.W.; supervision, Y.C., T.Z. and Y.H.; visualization, Y.J.; writing—original draft preparation, J.W.; writing—review and editing, W.Z. and Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Jilin Provincial Development and Reform Commission's special project for innovation ability construction (infrared image super-resolution and detail enhancement system research and development).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhu, D.; Zhan, W.; Jiang, Y. MIFFuse: A multi-level feature fusion network for infrared and visible images. *IEEE Access* **2021**, *9*, 130778–130792. [[CrossRef](#)]
2. Zhu, D.; Zhan, W.; Fu, J. RI-MFM: A Novel Infrared and Visible Image Registration with Rotation Invariance and Multilevel Feature Matching. *Electronics* **2022**, *11*, 2866. [[CrossRef](#)]
3. Qiu, S.; Luo, J.; Yang, S. A moving target extraction algorithm based on the fusion of infrared and visible images. *Infrared Phys. Technol.* **2019**, *98*, 285–291. [[CrossRef](#)]
4. Zhang, A.; Min, Z.; Zhang, Z. Generalized Point Set Registration with Fuzzy Correspondences Based on Variational Bayesian Inference. *IEEE Trans. Fuzzy Syst.* **2022**, *30*, 1529–1540. [[CrossRef](#)]
5. Yang, Z.; Zeng, S. TPFusion: Texture preserving fusion of infrared and visible images via dense networks. *Entropy* **2022**, *24*, 294. [[CrossRef](#)] [[PubMed](#)]
6. Zhu, D.; Zhan, W.; Jiang, Y. IPLF: A Novel Image Pair Learning Fusion Network for Infrared and Visible Image. *IEEE Sens. J.* **2022**, *22*, 8808–8817. [[CrossRef](#)]
7. Xu, H.; Ma, J.; Jiang, J. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [[CrossRef](#)]
8. Zhang, Y.; Liu, Y.; Sun, P. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [[CrossRef](#)]
9. Long, Y.; Jia, H.; Zhong, Y. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Inf. Fusion* **2021**, *69*, 128–141. [[CrossRef](#)]
10. Ma, J.; Zhang, H.; Shao, Z. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5005014. [[CrossRef](#)]
11. Ma, J.; Tang, L.; Xu, M. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5009513. [[CrossRef](#)]
12. Xu, D.; Wang, Y.; Xu, S. Infrared and visible image fusion with a generative adversarial network and a residual network. *Appl. Sci.* **2020**, *10*, 554. [[CrossRef](#)]
13. Xu, D.; Wang, Y.; Zhang, X. Infrared and visible image fusion using a deep unsupervised framework with perceptual loss. *IEEE Access* **2020**, *8*, 206445–206458. [[CrossRef](#)]
14. Tang, L.; Yuan, J.; Zhang, H. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83*, 79–92. [[CrossRef](#)]
15. Ma, J.; Yu, W.; Liang, P. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [[CrossRef](#)]
16. Ma, J.; Xu, H.; Jiang, J. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [[CrossRef](#)]
17. Hou, J.; Zhang, D.; Wu, W. A generative adversarial network for infrared and visible image fusion based on semantic segmentation. *Entropy* **2021**, *23*, 376. [[CrossRef](#)]
18. 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](#)]
19. 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](#)]
20. Guo, M.H.; Xu, T.X.; Liu, J.J. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [[CrossRef](#)]
21. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [[CrossRef](#)]
22. Wang, L.; Yao, W.; Chen, C. Driving behavior recognition algorithm combining attention mechanism and lightweight network. *Entropy* **2022**, *24*, 984. [[CrossRef](#)] [[PubMed](#)]
23. Hui, Y.; Wang, J.; Shi, Y. Low Light Image Enhancement Algorithm Based on Detail Prediction and Attention Mechanism. *Entropy* **2022**, *24*, 815. [[CrossRef](#)] [[PubMed](#)]

24. Li, H.; Qiu, K.; Chen, L. SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 905–909. [[CrossRef](#)]
25. Tao, H.; Geng, L.; Shan, S. Multi-Stream Convolution-Recurrent Neural Networks Based on Attention Mechanism Fusion for Speech Emotion Recognition. *Entropy* **2022**, *24*, 1025. [[CrossRef](#)] [[PubMed](#)]
26. Li, J.; Huo, H.; Li, C. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans. Multimed.* **2020**, *23*, 1383–1396. [[CrossRef](#)]
27. Li, J.; Li, B.; Jiang, Y. MSAt-GAN: A generative adversarial network based on multi-scale and deep attention mechanism for infrared and visible light image fusion. *Complex Intell. Syst.* **2022**, *8*, 4753–4781. [[CrossRef](#)]
28. Wang, X.; Girshick, R.; Gupta, A. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
29. Mishra, S.; Liang, P.; Czajka, A. CC-NET: Image complexity guided network compression for biomedical image segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; IEEE: Manhattan, NY, USA, 2019; pp. 57–60.
30. Zhu, Z.; Xu, M.; Bai, S. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 593–602.
31. Piao, J.; Chen, Y.; Shin, H. A new deep learning based multi-spectral image fusion method. *Entropy* **2019**, *21*, 570. [[CrossRef](#)]
32. He, K.; Gkioxari, G.; Dollár, P. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
33. Ma, J.; Chen, C.; Li, C.; Huang, J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion* **2016**, *31*, 100–109. [[CrossRef](#)]
34. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [[CrossRef](#)] [[PubMed](#)]
35. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522.
36. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 313–315. [[CrossRef](#)]
37. Rao, Y.J. In-fibre Bragg grating sensors. *Meas. Sci. Technol.* **1997**, *8*, 355. [[CrossRef](#)]
38. Eskicioglu, A.M.; Fisher, P.S. Image quality measures and their performance. *IEEE Trans. Commun.* **1995**, *43*, 2959–2965. [[CrossRef](#)]
39. 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](#)]
40. Xydeas, C.S.; Petrovic, V. Objective image fusion performance measure. *Electron. Lett.* **2000**, *36*, 308–309. [[CrossRef](#)]
41. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
42. Wang, Q.; Wu, B.; Zhu, P. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; IEEE: Manhattan, NY, USA, 2020; pp. 13–19.