CDFFusion: A Color-Deviation-Free Fusion Network for Nighttime Infrared and Visible Images

Hao Chen , Ting-Hua Zhang * , Shi-Jie Zhai , Xiao-Yun Tong , Rui Zhu

Posted Date: 14 October 2025

doi: 10.20944/preprints202510.0998.v1

Keywords: night scene; image enhancement; image fusion

Article

CDFFusion: A Color-Deviation-Free Fusion Network for Nighttime Infrared and Visible Images

Hao Chen, Ting-Hua Zhang *, Shi-Jie Zhai, Xiao-Yun Tong and Rui Zhu

Space Engineering University, Beijing, 101416, China

* Correspondence: zth-gd@163.com; Tel.: +86-186-1298-0800

Abstract

The purpose of infrared and visible image fusion is to integrate their complementary information into a single image, thereby increasing the amount of information expression. However, previous methods often struggle to extract information hidden in darkness, and existing methods, which integrate brightness enhancement and image fusion, can cause overexposure, image blocking effects, and color deviation. Therefore, we propose a visible light and infrared image fusion method, CDFFusion, for low-light scenarios. Specifically, our method consists of two stages: First, an encoder is designed to extract deep features of visible light and infrared images respectively. Then, combined with RetiNex theory, a decomposition network is designed at the feature level to separate the illuminance component and reflectance component of the visible light image. Next, the proposed formula is used to process the Cb and Cr components of the original visible light image. The features of the reflectance component and infrared features are concatenated and input into the fusion network to obtain the Y component of the fused image. Finally, it is concatenated with the processed Cb' and Cr' components of the visible light image to get the final fused image. Experimental results show that the proposed method can effectively alleviate overexposure and image blocking effects, and there is no color deviation at all.

Keywords: night scene; image enhancement; image fusion

1. Introduction

Visible images usually contain rich texture detail information, but they are prone to loss of target information in complex scenes; infrared images, which are formed based on thermal radiation information, are not easily affected by harsh conditions, but they lack detailed descriptions of the scene. Therefore, infrared and visible image fusion (IVIF) can make full use of their complementary information and significantly improve the comprehensive perception ability of the scene. However, in current low-light environments with low visibility, most of the texture details in visible light images are obscured. The general approach to address this issue is to first preprocess the visible light image using a low-light enhancement method and then fuse it with the infrared image. However, this can cause color distortion in some areas, so how to organically combine low-light image enhancement with IVIF is a significant challenge.

Figure 1 shows the visible image, infrared image, the result processed by RFN-Net, and the result processed by LEDNet [1] preprocessing followed by RFN-Net. Firstly, previous IVIF methods fail to extract the information of visible images obscured at night (as shown in the red box in (c)). In contrast, fusing the visible image after enhancement preprocessing causes color distortion in some areas (as shown in the green box in (d), where white zebra crossings are rendered green).

Secondly, existing nighttime IVIF methods, while solving the above two problems, introduce new issues. Figure 2 displays the visible image, infrared image, and the results processed by DIVFusion, LEFuse, LENFusion, and the proposed method. It can be observed that DIVFusion and LEFuse overemphasize the difference between high and low gray values and focus on highlighting high-gray-value regions, which leads to two consequences: first, overexposure occurs in certain parts

 doi:10.20944/preprints202510.0998.v1

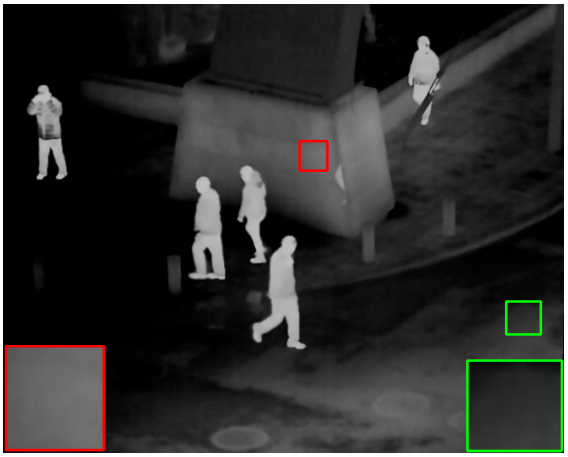
of the image (the lower left corners in (c) and (d)); second, severe block artifacts (mosaic effect, as shown in the red boxes in (c) and (d)) or false edges appear in non-edge or weak-edge regions. Although LENFusion avoids overexposure, it differs from the previous two methods by focusing on suppressing low-gray-value regions, which also results in two problems: in some weak-edge regions, the low-gray-value parts are severely distorted, leading to information loss (as shown in the red box in (e)); in other weak-edge regions, the colors of these low-gray-value parts are rendered extremely dark, causing unnatural false edges (not shown in the figure). Additionally, these methods only process the Y channel of visible images without handling the Cb and Cr channels, which changes the hue and saturation of the source images, resulting in varying degrees of color deviation.

To solve the above problems, this paper proposes a two-stage network for joint low-light image enhancement and image fusion without color deviation. It can alleviate overexposure and block artifacts while eliminating color deviation. Firstly, an encoder with a Feature Pyramid Network (FPN) structure is used to extract deep features of visible and infrared images respectively. Then, combined with the RetiNex theory, a decomposition network is designed at the feature level to separate the illumination component and reflectance component of the visible image. Next, the proposed formula is applied to process the Cb and Cr components of the original visible image. The reflectance component features and infrared features are concatenated and input into the fusion network to obtain the Y component of the fused image. Finally, the Y component is concatenated with the processed Cb' and Cr' components of the visible image to generate the final fused image.

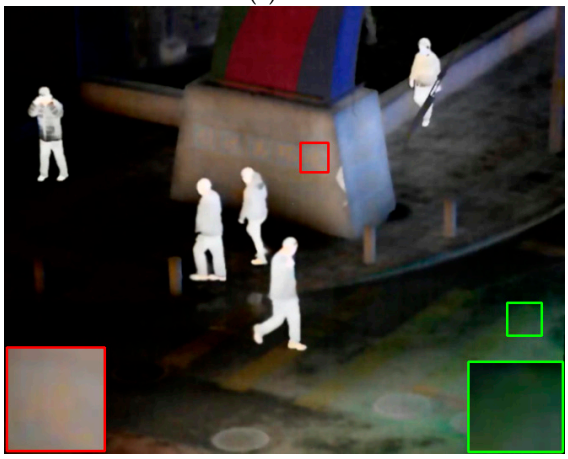
In summary, the main contributions of this paper are as follows:

- This paper proposes a two-stage network for joint low-light image enhancement and image fusion, which can alleviate overexposure and block artifacts while eliminating color deviation, named CDFFusion;
- A brightness enhancement formula without color deviation is proposed, which processes the three components (Y, Cb, Cr) simultaneously, and the processed results have no color deviation;

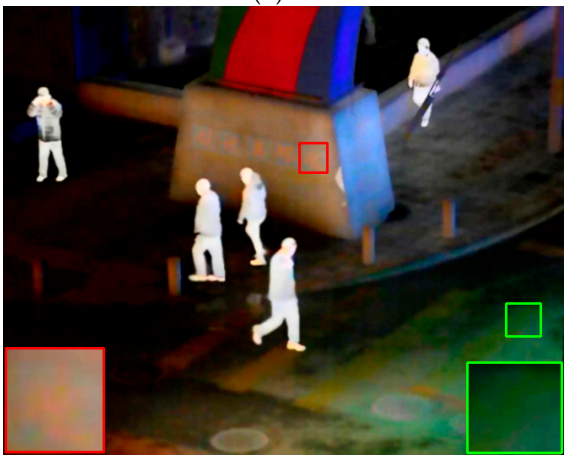
(**a**) VIS



(**b**) IR



(**c**) RFN-Nest



(**d**) LEDNet+RFN-Nest

 doi:10.20944/preprints202510.0998.v1

Figure 1. Fusion results of previous methods on the LLVIP dataset.

(**a**) VIS (**b**) IR (**c**) DIVFusion

(**d**) LEFuse (**e**) LENFusion (**f**) ours

Figure 2. Fusion results of some recent methods and the proposed method on the LLVIP dataset.

2. Related Work

Deep learning-based visible and infrared image fusion methods can be roughly divided into three categories: those based on Convolutional Neural Networks (CNN), those based on Autoencoders (AE), and those based on Generative Adversarial Networks (GAN). Zhang et al. proposed a Squeeze-and-Decomposition Network, SD-Net [2], which models image fusion as the extraction and reconstruction of gradient and intensity information. Based on the fact that different fusion tasks share a similar goal—fusing images by integrating important and complementary information from multiple source images—Xu et al. proposed a unified unsupervised end-to-end image fusion network, U2Fusion [3], which is characterized by its ability to perform information measurement on extracted features to automatically estimate the importance of source images. Li et al. proposed an end-to-end fusion network architecture, RFN-Nest [4], which includes an encoder, a Residual Fusion Network (RFN), and a decoder. The encoder extracts multi-scale deep features through max-pooling; the RFN is composed of several convolutional layers and is trained with a new loss function to implement a learnable fusion strategy instead of manually designed rules; the decoder adopts an architecture based on nested connections to reconstruct the fused image. Xu et al. proposed a classification saliency-based pixel-level fusion method, CSF [5], which classifies different source images, uses the classification results to represent the saliency of each pixel in the source images, and finally fuses the feature maps using this saliency to generate the fusion result. Ma et al. first proposed an end-to-end model based on Generative Adversarial Networks (GAN) and named it FusionGAN [6]. It formulates the fusion problem as an adversarial problem while avoiding the manual design of complex activity level measurement and fusion rules. Later, the team designed another Generative Adversarial Network with Multiclassification Constraints, GANMcC [7], which formulates the fusion problem as the simultaneous estimation of multiple distributions.

DIVFusion [8] was the first to combine IVIF and low-light enhancement. It decomposes the visible image into reflectance features and illumination features at the feature level, then enhances and fuses the reflectance features with infrared features. LENFusion [9] adopts the idea of pre-enhancement, re-enhancement, and fusion. It draws on the idea of PIAFusion [10] by pre-training a binary classifier and using its classification results as part of the loss function of the backbone

 doi:10.20944/preprints202510.0998.v1

network. LEFuse [11] introduces a hybrid module of Transformer and CNN and designs the overall network structure into a symmetric structure similar to the U-net style.

3. Methods

The proposed method in this paper improves nighttime visibility, retains the complementary information of the two modalities, alleviates overexposure and block artifacts, and eliminates color deviation. This section details the two sub-networks of the entire framework, including the network structure and loss function.

3.1. Overall Framework

All As shown in Figure 3, let I_{vi}^{Y} and I_{ir} represent the Y channel of the visible image and the infrared image respectively. After passing through the decomposition network and fusion network, the Y channel of the fused image is obtained. Then, it is concatenated with the processed Cb and Cr channels of the visible image to get the final fusion result. The entire process is divided into two stages: feature extraction and image fusion.

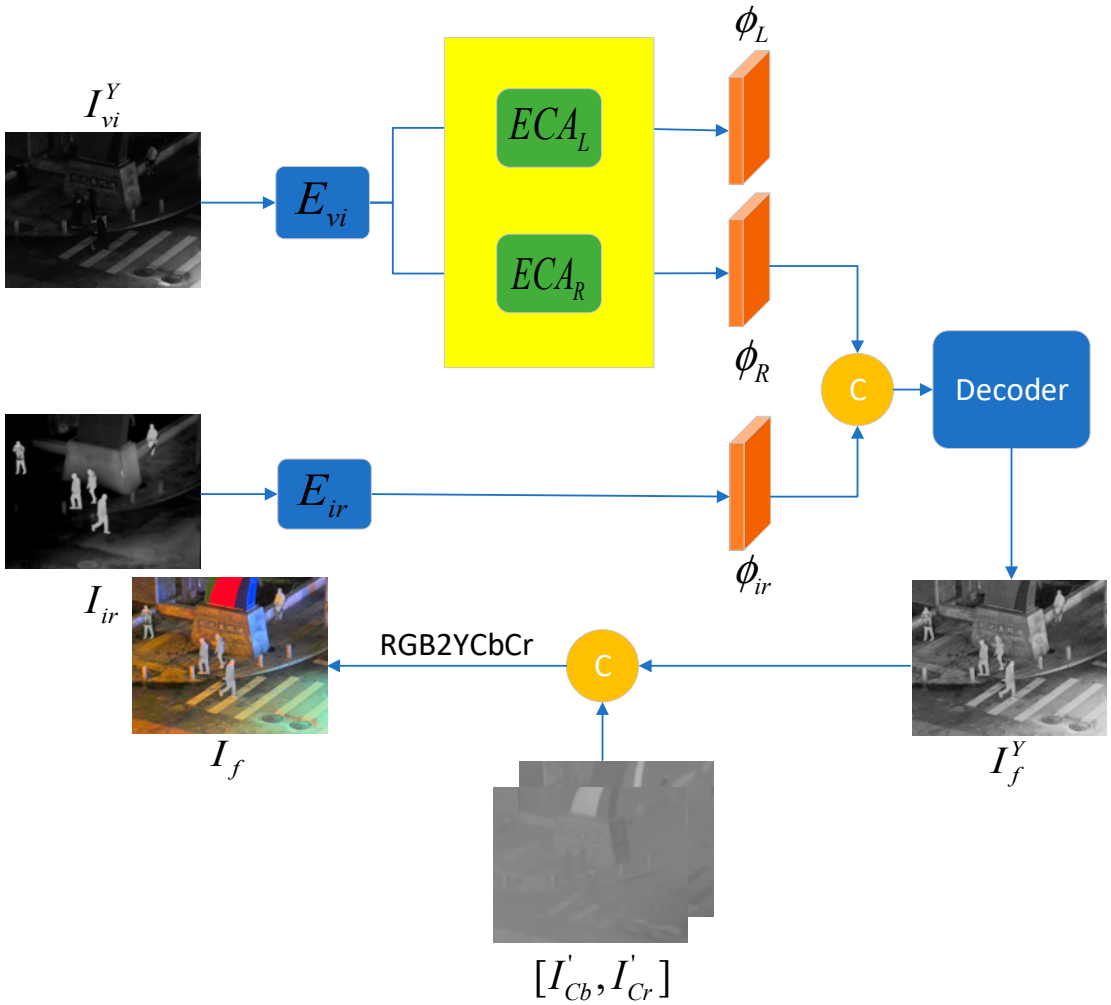


Figure 3. This Overall framework of CDFFusion.

 doi:10.20944/preprints202510.0998.v1

3.2. Reflectance-Illumination Decomposition Network(RID-Net)

The specific structure of the reflectance-illumination decomposition network is shown in Figure 4, which consists of two parts: a visible image reconstruction network and an infrared image reconstruction network. The former is used to separate the reflectance and illumination components of the visible image at the feature level and reconstruct them into images, while the latter is used to extract deep features of the infrared image and reconstruct them into images. First, in the feature extraction stage, the Y channel of the visible image is input into the encoder with an FPN structure to extract deep features ϕ. The specific structure of the encoder is shown in Figure 5. The original image passes through 4 convolutional layers (C1-C4) in sequence to obtain 4 feature maps of different sizes. The parameters of C1-C4 are listed in Table 1. Then, 4 1×1 convolutional layers (C5-C8) are used to unify the number of channels of the feature maps obtained just now, which is set to 256 here. After that, these processed feature maps are upsampled by 2 times and summed to complete multi-scale fusion. This process can be expressed as:

$$\phi = E(I_{vi}^{Y}) \tag{1}$$

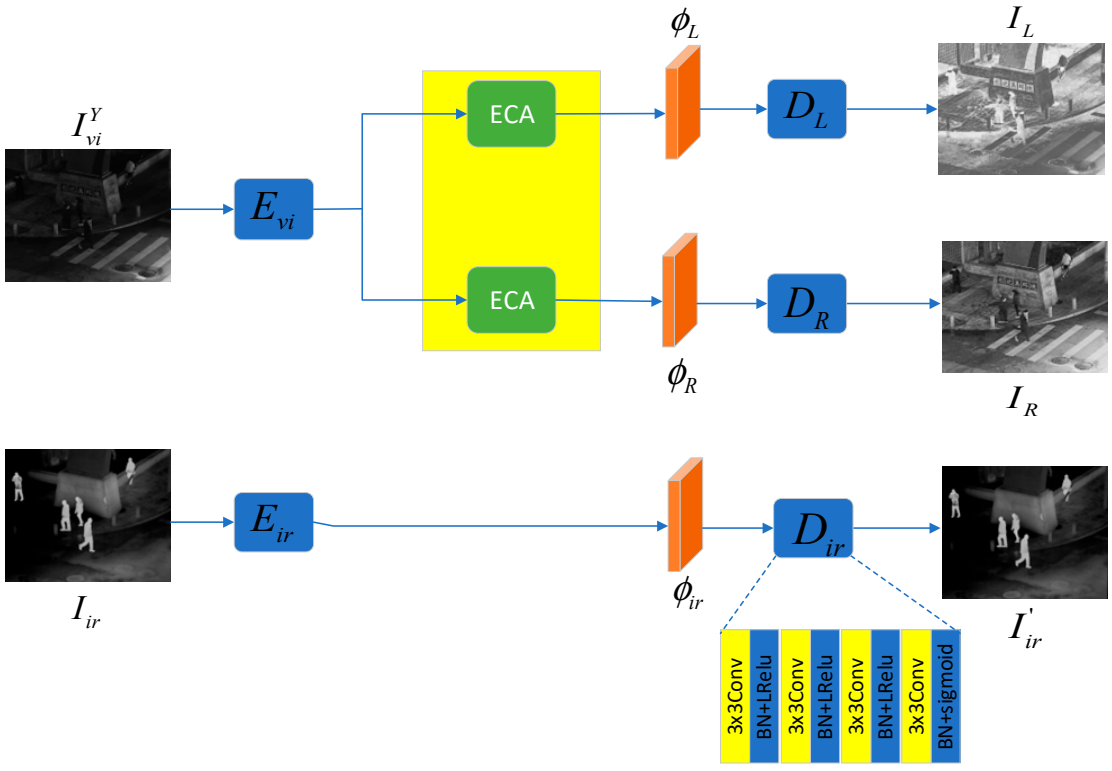


Figure 4. Structure of the RID-Net.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 October 2025 doi:10.20944/preprints202510.0998.v1

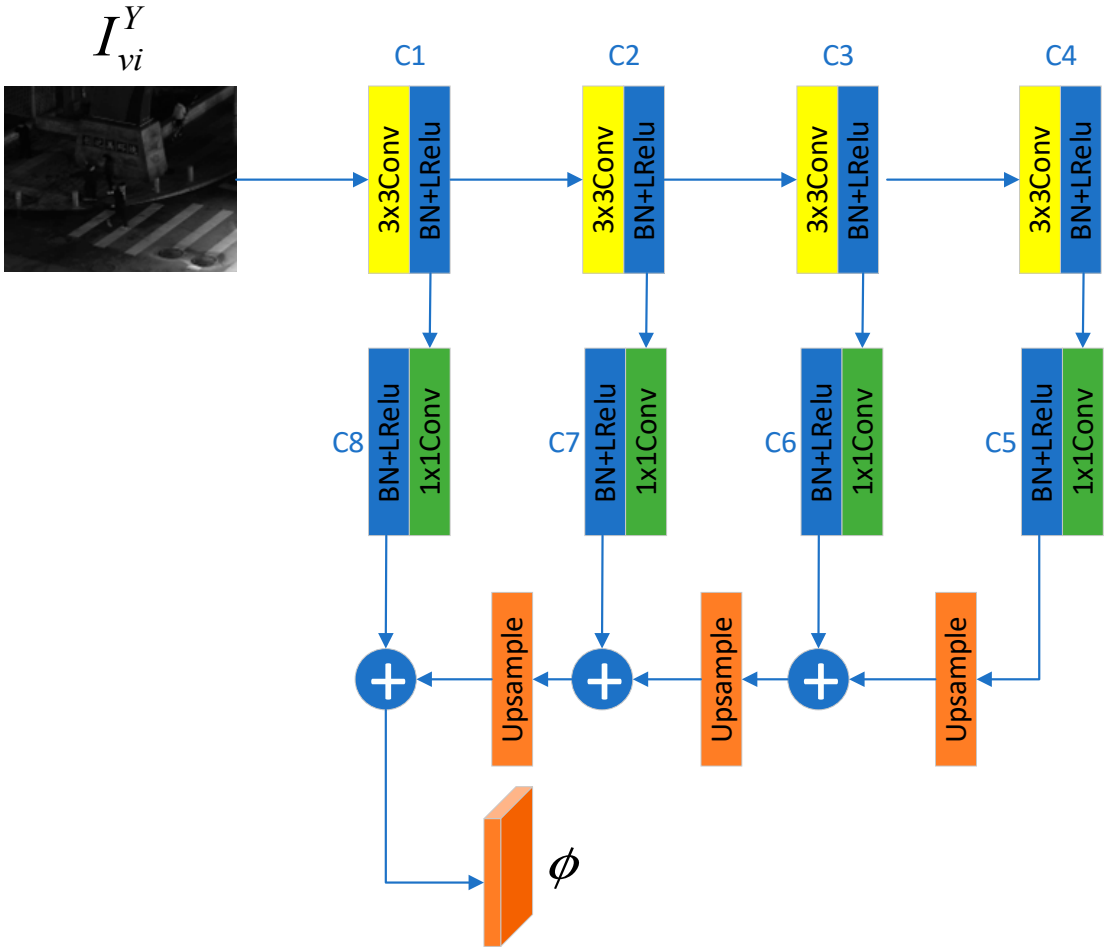


Figure 5. Structure of the encoder.

Table 1. Parameters of C1-C4.

	In channels	Out channels	Kernel size	stride	padding
C1	1	64	3	1	1
C2	64	128	3	2	1
C3	128	256	3	2	1
C4	256	512	3	2	1

In the feature separation stage, according to the RetiNex theory, an image can be decomposed into the product of the reflectance component (R) and the illumination component (L) (as shown in the following formula). Therefore, two Efficient Channel Attention (ECA) modules are used to separate the reflectance features ϕ_R and illumination features ϕ_L from them.

$$I = R \times L \tag{2}$$

The structure of the ECA module is shown in Figure 6. First, global average pooling is performed on each channel of the input feature map to obtain the global feature of each channel. Then, one-dimensional convolution and activation are applied to these global features to obtain attention weights. The input feature map is reweighted using these weights to get the output feature. Here, GAP and σ represent global average pooling and sigmoid activation function respectively. This process can be expressed as:

 doi:10.20944/preprints202510.0998.v1

$$ECA_R(\phi) = \phi_R$$
$$ECA_L(\phi) = \phi_L \tag{3}$$

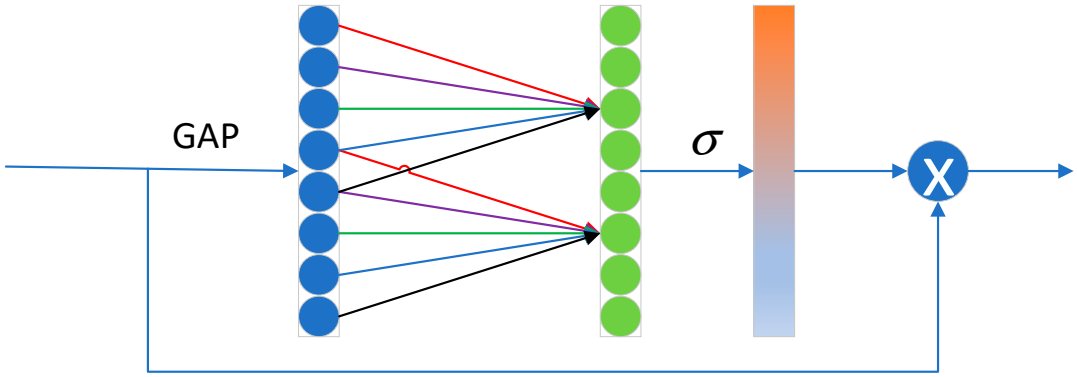


Figure 6. Structure of the ECA module.

In the final decoding stage, they are input into two decoders with identical structures, and the reflectance component I_R and illumination component I_L are obtained through reconstruction. Both decoders are composed of 4 stacked groups of convolutional and activation layers. The first 3 activation layers use the LRelu function, and the last one uses the sigmoid function. This process can be expressed as:

$$D_R(\phi_R) = I_R$$
$$D_L(\phi_L) = I_L \tag{4}$$

It should be noted that the decoding part here is only for better image reconstruction and does not participate in the operation of the next stage.

Similarly, the infrared reconstruction network adopts a basically consistent structure, which will not be elaborated here.

The loss function of this part of the network is as follows:

$$L_1 = \lambda_1 L_{rev} + \lambda_2 L_{per} \tag{5}$$

The loss function L_1 of the visible image reconstruction network consists of two parts: reconstruction loss L_{rev} and perceptual loss L_{per}. The reconstruction loss L_{rev} is derived from Formula (2), which is expressed as follows:

$$L_{rev} = \left\| I_R \times I_L - I_{vi}^Y \right\|_1 \tag{6}$$

On this basis, we hope that the brightness-enhanced result (i.e., the reflectance component I_R) and the result of the original image I_{vi}^Y after histogram equalization have as similar representations as possible in the feature domain of the VGG-19 network. Thus, the perceptual loss L_{per} is defined as:

$$L_{per} = \left\| VGG(I_R) - VGG[hist(I_{vi}^Y)] \right\|_2^2 \tag{7}$$

where "hist" represents the histogram equalization operation. Here, the Conv4-1 feature of the VGG-19 network is selected.

 doi:10.20944/preprints202510.0998.v1

The loss function L_2 of the infrared reconstruction network consists of reconstruction loss L_{rei} and structural similarity loss L_{s1}, which can reconstruct results similar to the original image in both intensity and structure. It is defined as follows:

$$\begin{aligned} L_2 &= \lambda_3 L_{rei} + \lambda_4 L_{s1} \\ L_{rei} &= \left\| I_{ir}' - I_{ir} \right\|_1 \\ L_{s1} &= 1 - ssim(I_{ir}', I_{ir}) \end{aligned} \tag{8}$$

Here, $\lambda_1 \sim \lambda_4$ are weight hyperparameters used to balance the various parts of the loss.

3.3. Fusion Network

In the fusion stage, the reflectance features ϕ_R of the visible image and the infrared features ϕ_{ir} are concatenated at the channel level and input into the fusion network. Consistent with the decoder structure in the previous stage, it is also composed of 4 stacked groups of convolutional and activation layers. At this time, its output I_f^Y is taken as the Y channel of the fused image, which is used to participate in the final channel fusion.

According to the **ITU-R BT.601** international standard, the conversion formula from the RGB space to the YCbCr space of an image is:

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ Cb &= 128 - 0.168736R - 0.331264G + 0.5B \\ Cr &= 128 + 0.5R - 0.418688G - 0.081312B \end{aligned} \tag{9}$$

In the HSI color space, the conversion formula from RGB to HSI is:

$$\begin{aligned} H &= \begin{cases} \theta, B \le G \\ 360^\circ - \theta, B > G \end{cases} \\ \theta &= \arccos\left\{ \frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2 + (R-B)(G-B)^{1/2}]} \right\} \\ S &= 1 - \frac{3}{(R+G+B)}[\min(R,G,B)] \end{aligned} \tag{10}$$

where H represents hue and S represents saturation. From the above formula, it can be proved that: only when the three components R, G, and B are scaled proportionally, the hue and saturation of the pixel remain unchanged. Substituting the result into Formula (9), it can be obtained that: only when (Cb - 0.5) and (Cr - 0.5) are scaled synchronously with Y, R, G, and B are scaled proportionally (the pixel value range has been normalized to [0, 1]), and the hue and saturation remain unchanged. Thus, we naturally derive the mapping formula for the Cb and Cr channels of the visible image:

$$\begin{aligned} scale &= I_R / I_{vi}^Y \\ I_{vi}^{Cb}' - 0.5 &= (I_{vi}^{Cb} - 0.5) \times scale \\ I_{vi}^{Cr}' - 0.5 &= (I_{vi}^{Cr} - 0.5) \times scale \end{aligned} \tag{11}$$

where "scale" is the brightness gain image. Different from other methods that use loss functions for weak constraints, the above formula applies strong constraints on the proportion between the three components R, G, and B of the visible image, thus avoiding color distortion. Finally, I_f^Y is

 doi:10.20944/preprints202510.0998.v1

concatenated with I_{vi}^{Cb}' and I_{vi}^{Cr}', and converted back to the RGB space to obtain the final fused image I_f.

The loss function L_3 of the fusion network is as follows:

$$L_3 = \lambda_5 L_{vi} + \lambda_6 L_{ir} + \lambda_7 L_{aux} + \lambda_8 L_{grad} + \lambda_9 L_{s2} \tag{12}$$

It consists of visible light intensity loss L_{vi}, infrared intensity loss L_{ir}, auxiliary intensity loss L_{aux}, gradient loss L_{grad}, and structural similarity loss L_{s2}, which are defined as follows:

$$\begin{aligned} L_{vi} &= \left\| I_f^Y - I_R \right\|_1 \\ L_{ir} &= \left\| I_f^Y - I_{ir} \right\|_1 \\ L_{aux} &= \left\| I_f^Y - \max(I_R, I_{ir}) \right\|_1 \\ L_{grad} &= \left\| \nabla I_f^Y - \max(\nabla I_R, \nabla I_{ir}) \right\| \\ L_{s2} &= 1 - \frac{1}{2}[ssim(I_f^Y, I_R) + ssim(I_f^Y, I_{ir})] \end{aligned} \tag{13}$$

where ∇ represents the gradient operation, and the Sobel operator is used here. These losses can force the fused image to retain more prominent intensity and gradient information from the source images and maintain a relatively consistent structural similarity with the source images. Here, $\lambda_5 \sim \lambda_9$ are weight hyperparameters used to balance the various parts of the loss.

4. Experiments

4.1. Experimental Configuration

To comprehensively evaluate the proposed method, extensive experiments are conducted on the LLVIP dataset [12]. The LLVIP dataset is a paired visible-infrared dataset for low-light scenarios, containing 33,672 images (16,836 pairs). Among them, 240 pairs of infrared and visible images are selected for the training phase, and 50 pairs are selected for the testing phase. These images have been strictly registered. The results of this paper are compared with 5 fusion methods, including 1 AE-based method (RFN-Nest), 1 GAN-based method (FusionGAN), and three latest nighttime IVIF methods (DIVFusion, LEFuse, and LENFusion). The implementation of all methods is based on publicly available code.

In the quantitative evaluation phase, six metrics are used, including 1 image feature-based metric (Spatial Frequency, SF), 1 structural similarity-based metric (Multi-Scale Structural Similarity, MS-SSIM), and 4 metrics based on the source images and generated images (Correlation Coefficient, CC; Sum of Correlation Differences, SCD; Gradient-Based Fusion Performance, Qabf; and Noise-Based Fusion Performance, Nabf). Among them, a smaller Nabf value indicates a better fusion effect.

In the training phase, the 240 pairs of images used for training are randomly cropped to a size of 224×224 pixels, and the batch size is set to 5. The initial learning rates of the visible and infrared images in the decomposition network are 10^-4 and 10^-3 respectively, and they are decayed to 0.1 and 0.01 of the initial values after 50 and 75 epochs. The fusion network uses a fixed learning rate set to 2×10^-5. In addition, the weight hyperparameters of the loss functions of the decomposition network and fusion network are set as follows:

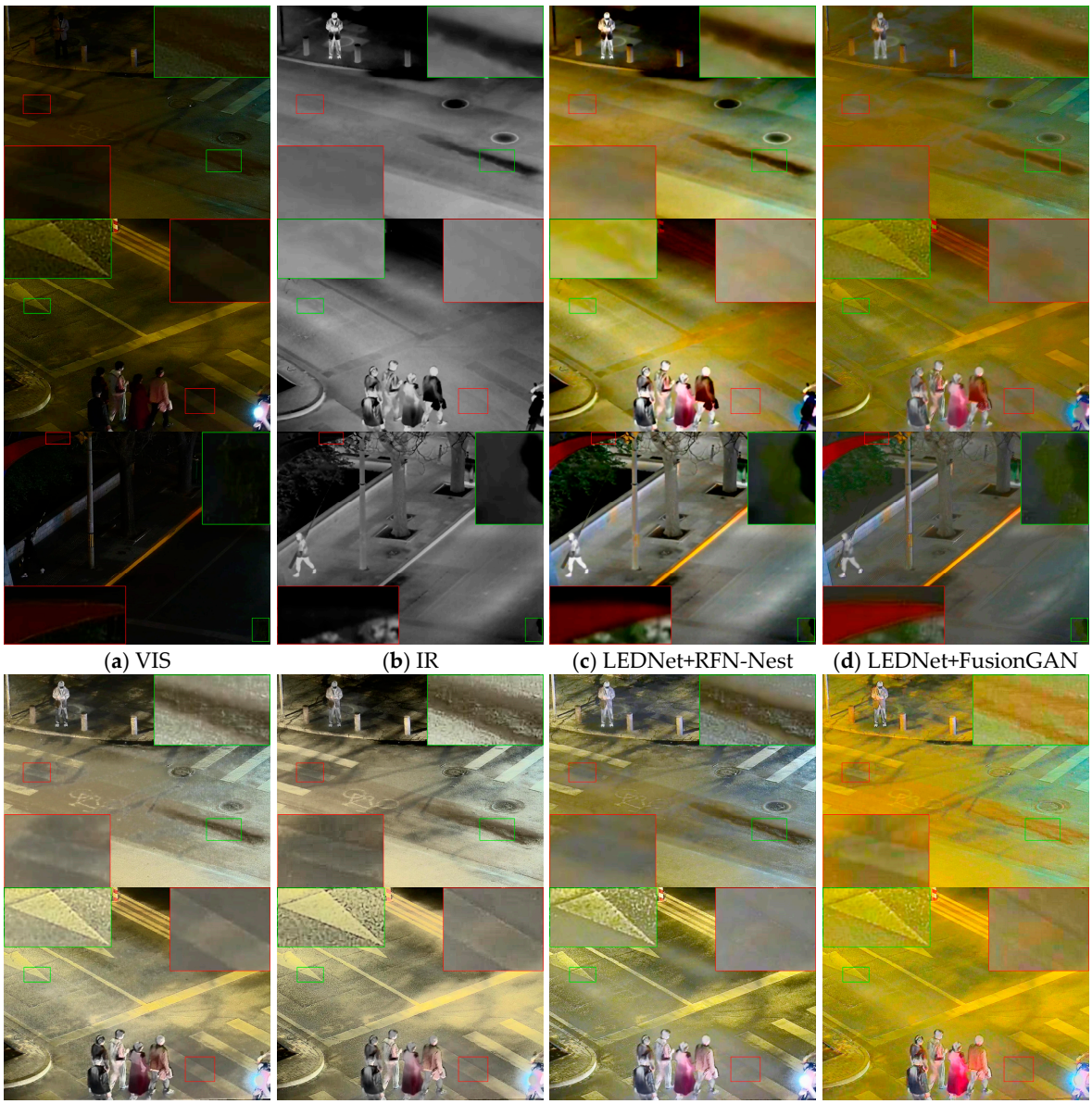
$$\lambda_1 = 10,\ \lambda_2 = 5,\ \lambda_3 = 10,\ \lambda_4 = 1,\ \lambda_5 = 2, \lambda_6 = 2,\ \lambda_7 = 6,\ \lambda_8 = 10,\ \lambda_9 = 30 \tag{14}$$

 doi:10.20944/preprints202510.0998.v1

4.2. Results and Analysis

4.2.1. Qualitative Comparison

Figure 7 shows the visualization results of different fusion methods. As mentioned earlier, previous fusion methods fail to extract the information of visible images obscured at night. In the results of RFN-Nest and FusionGAN, the upper part of the images in the first row is still completely black, with no significant improvement in brightness, and the zebra crossings in the red boxes become blurred, with the zebra crossings in the upper right corner showing a greenish tint. In the second row of images, most details of the road in the boxes are lost, and the headlight area in the lower right corner is covered with a shadow. In the third row of images, the road details are completely lost. For the results of DIVFusion, LEFuse, and LENFusion, the color saturation in figures (e) and (f) is relatively low, leading to an overexposed feeling in some areas (as shown in the green boxes in the second row of images). In figure (g), areas such as zebra crossings are prone to distortion, resulting in information loss (as shown in the red boxes in the first and second rows of images). In areas such as damaged roads, the color is darkened, forming unnatural false edges (as shown in the green box in the first row of images). In addition, these three methods all cause block artifacts (as shown in the boxes in the third row of images). In the third row of images, the color saturation of buildings and leaves is significantly low, with varying degrees of color deviation. In contrast, the proposed method alleviates all the above issues and achieves a better visual effect without color deviation.



(**a**) VIS (**b**) IR (**c**) LEDNet+RFN-Nest (**d**) LEDNet+FusionGAN

 doi:10.20944/preprints202510.0998.v1

(**e**) DIVFusion (**f**) LEFuse (**g**) LENFusion (**h**) ours

Figure 7. Experimental results of different methods on the LLVIP dataset.

4.2.2. Quantitative Comparison

Table 2 shows the average quality metrics of different fusion methods on the LLVIP dataset. It can be seen that the proposed method ranks first in the Qabf and MS-SSIM metrics, second in the CC and SCD metrics, and third in the SF and Nabf metrics. This indicates that the proposed method can suppress block artifacts while maintaining good high-frequency details, and can well maintain the consistency between the fused image and the source images.

Table 2. Quantitative comparison between the proposed method and other methods on the LLVIP dataset. The best and second-best results are marked in bold and underlined respectively.

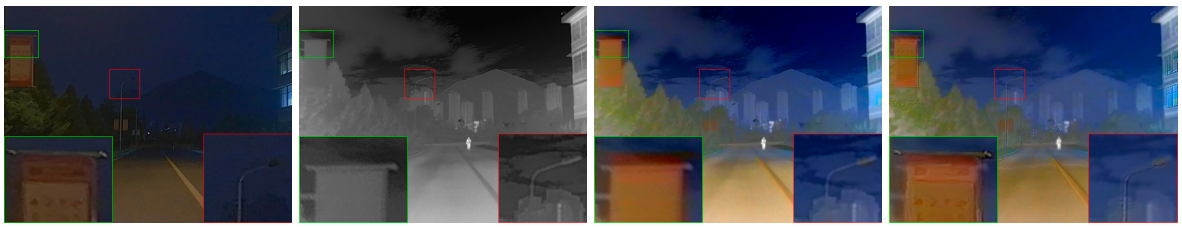
	SF	CC	Nabf	Qabf	SCD	MS-SSIM
RFN-Nest	5.0344	0.5824	**0.0099**	0.3332	1.0923	0.7730
FusionGAN	6.7466	0.6479	<u>0.0124</u>	0.2788	0.9159	<u>0.8188</u>
DIVFusion	14.7371	**0.6922**	0.1364	<u>0.3823</u>	**1.5373**	0.7973
LEFuse	**24.0328**	0.6087	0.1856	0.3006	1.2262	0.7027
LENFusion	<u>21.4990</u>	0.5928	0.1969	0.3534	1.0440	0.7236
ours	17.6531	<u>0.6619</u>	0.1075	**0.4279**	<u>1.2760</u>	**0.8335**

4.3. Generalization Experiment

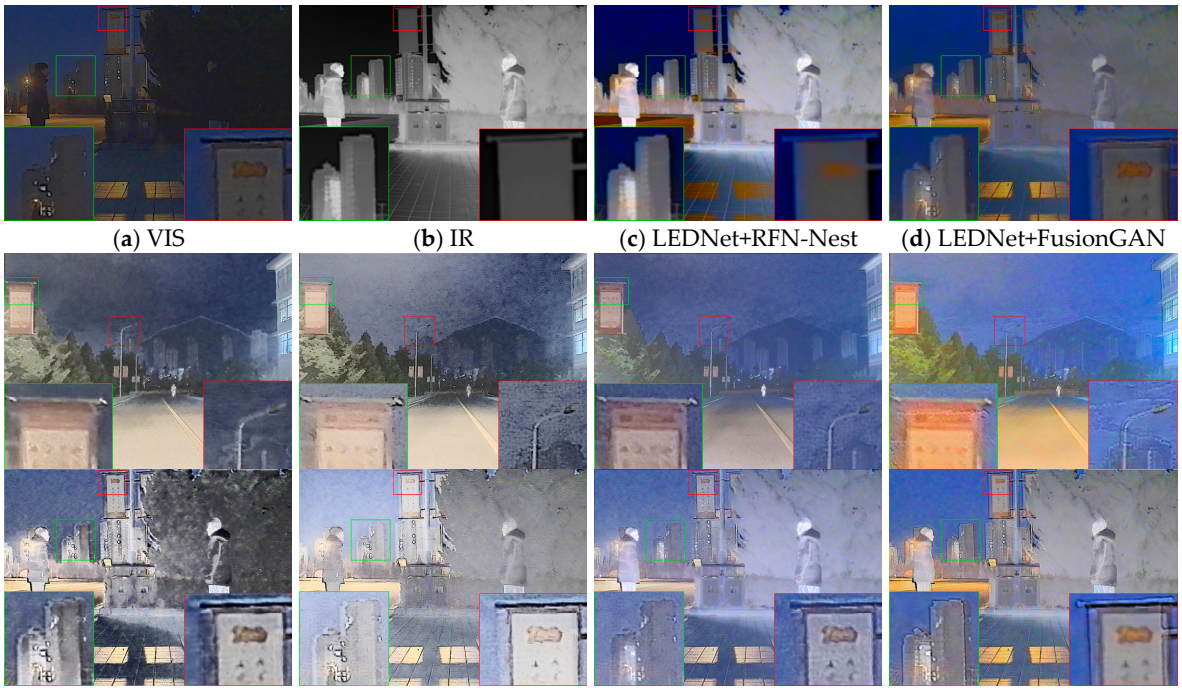
To verify the generalization ability of the proposed model, 48 pairs of images are selected from the M3FD dataset [13] for testing. The model is not trained on the M3FD dataset and is directly used for testing. The M3FD dataset is a visible-infrared fusion dataset, containing 4,200 pairs of images for fusion, detection, and fusion-based detection, as well as 300 pairs of images for independent scene fusion.

4.3.1. Qualitative Comparison

As shown in Figure 8, first, in the result of RFN-Nest, the information on the signs is obscured in darkness and cannot be seen clearly (as shown in the green box in the first row of images and the red box in the second row of images). Secondly, the results of RFN-Nest and FusionGAN are generally blurrier than the original images (such as the tree areas in the first row of images). In the results of DIVFusion, LEFuse, and LENFusion, snowflake-like noise appears in the sky area of the first row of images, and the outline of the clouds is eroded (as shown in the red box in the first row of images). In the second row of images, the outline of distant buildings becomes incomplete compared with the original image. In addition, the results of these three methods all have varying degrees of color distortion, and figure (f) even gives an overexposed feeling.



 doi:10.20944/preprints202510.0998.v1



(**a**) VIS (**b**) IR (**c**) LEDNet+RFN-Nest (**d**) LEDNet+FusionGAN

(**e**) DIVFusion (**f**) LEFuse (**g**) LENFusion (**h**) ours

Figure 8. Experimental results of different methods on the M3FD dataset.

4.3.2. Quantitative Analysis

Table 3 shows the average quality metrics of different fusion methods on the M3FD dataset. It can be seen that the proposed method ranks first in the Qabf metric, second in the SF and CC metrics, and third in the Nabf, SCD, and MS-SSIM metrics. This indicates that the proposed method can retain rich texture details and well maintain the correlation and consistency between the fused image and the source images.

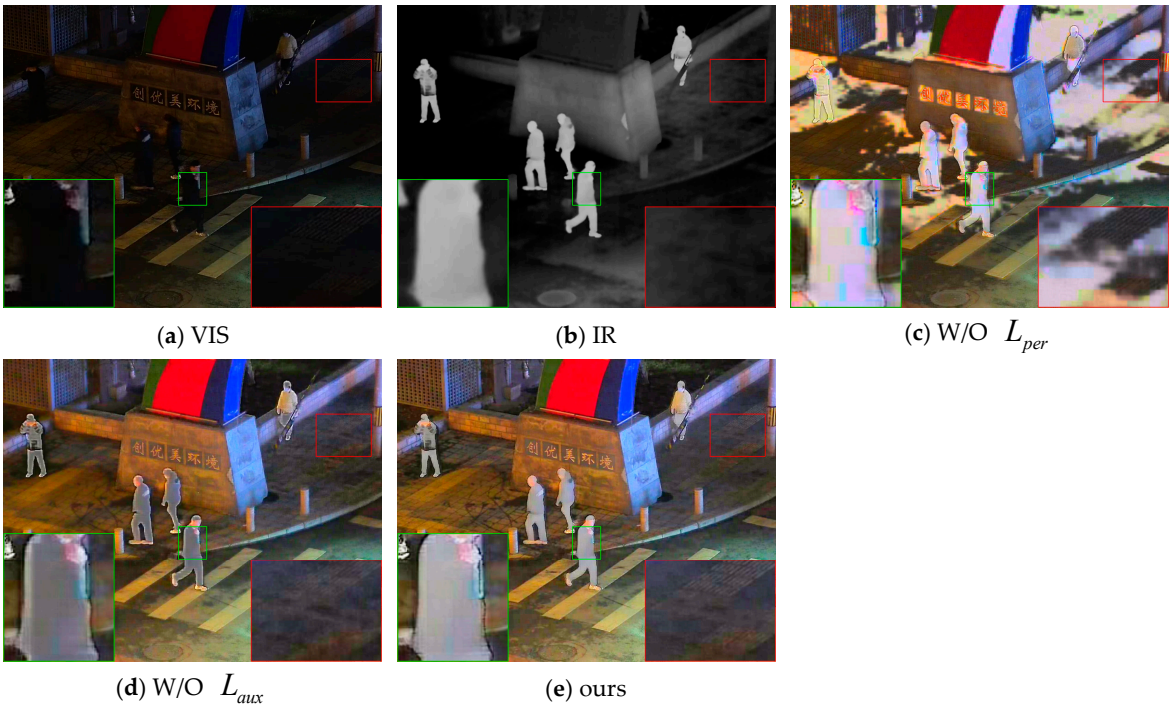
Table 3. Quantitative comparison between the proposed method and other methods on the M3FD dataset. The best and second-best results are marked in bold and underlined respectively.

	SF	CC	Nabf	Qabf	SCD	MS-SSIM
RFN-Nest	4.1495	0.6260	**0.0039**	0.3133	0.9690	<u>0.8565</u>
FusionGAN	6.2380	0.6579	<u>0.0143</u>	0.2770	0.7082	**0.8788**
DIVFusion	14.4927	**0.7360**	0.1525	<u>0.3684</u>	**1.6140**	0.8276
LEFuse	**20.6205**	0.6623	0.2416	0.2504	<u>1.2501</u>	0.7236
LENFusion	14.6193	0.6640	0.1402	0.3637	0.9860	0.8011
ours	<u>16.4501</u>	<u>0.6879</u>	0.1251	**0.3694**	1.0070	0.8416

4.4. Ablation Experiment

To verify the effectiveness of each loss function in the proposed method, an ablation experiment is designed. To verify the effectiveness of the perceptual loss L_{per} and auxiliary intensity loss L_{aux}, these two loss functions are removed respectively, and the results are shown in Figure 9.

 doi:10.20944/preprints202510.0998.v1



(**a**) VIS (**b**) IR (**c**) W/O L_{per}

(**d**) W/O L_{aux} (**e**) ours

Figure 9. Comparison of results of the ablation experiment on perceptual loss L_{per} and auxiliary intensity loss L_{aux} based on the LLVIP dataset.

4.4.1. Qualitative Comparison

When the perceptual loss L_{per} is removed, the fused image exhibits severe color deviation and block artifacts. In fact, at this time, the RID-Net cannot accurately separate the reflectance component I_R. When the auxiliary intensity loss L_{aux} is removed, the fused image fails to express sufficient infrared information (as shown in the green box in Figure (d)), and a lot of road texture information is lost (as shown in the red box in Figure (d)).

4.4.2. Quantitative Analysis

The quantitative results of the ablation experiment are shown in Table 4. It can be seen that the proposed method achieves better results in the CC, Qabf, and SCD metrics. The SF is not the optimal because areas with block artifacts often have higher spatial frequencies. The MS-SSIM is also not the optimal because without the auxiliary intensity loss L_{aux}, the fused image tends to be closer to either the visible image or the infrared image.

Table 4. Results of the ablation experiment based on the LLVIP dataset. The best results are marked in bold.

	SF	CC	Nabf	Qabf	SCD	MS-SSIM
W/O L_{per}	**17.7592**	0.209	**0.0868**	0.4103	0.1861	0.6925
W/O L_{aux}	15.3465	0.6607	0.0957	0.4169	1.242	**0.8699**
ours	17.6531	**0.6619**	0.1075	**0.4279**	**1.276**	0.8335

5. Conclusions and Discussion

This paper proposes a new fusion method to address the shortcomings of existing visible and infrared image fusion methods in low-light scenarios. CDFFusion adopts a two-stage strategy to solve the problems of overexposure and block artifacts, and completely avoids color deviation. Specifically, we first train a feature extraction network that can extract features of source images and separate the illumination component and reflectance component of the visible image. Then, a fusion

 doi:10.20944/preprints202510.0998.v1

network is trained, which combines the proposed color mapping formula to realize fusion of infrared images and enhanced visible images without color deviation. Experimental results show that compared with the existing five methods, the proposed method achieves generally better results in both subjective and objective aspects.

The limitations of this paper are as follows: the encoder with the FPN structure has certain requirements on the size of the input image, requiring both the height and width of the image to be integer multiples of 8. Images that do not meet the size requirements need to be preprocessed. In addition, there are certain shortcomings in the feature extraction of infrared images, and the unique advantages of infrared images are not separated and fully utilized.

Appendix A

Theorem 1. *Condition for Constant H and S.*

Two pixels have the same H and S if and only if the three terms (Y, Cb - 0.5, Cr - 0.5)(normalized) of the two pixels are proportional to each other respectively. follows:

Proof of Theorem 1. Let $x=(R,G,B)^T$ and $x^e=(R^e,G^e,B^e)^T$ .According to Equation (10), we have $\theta(x)=\frac{\frac{1}{2}px}{\sqrt{\frac{1}{2}x^TAx}}$ and $S(x)=1-\frac{3\min(x)}{(1,1,1)x}$, where $p=(2,-1,-1)$, $A=\begin{pmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{pmatrix}$, and x^TAx are non - zero.

Now, consider the following system of equations:

$$\begin{cases}H(x)=H(x^e)\\S(x)=S(x^e)\end{cases} \tag{15}$$

To further simplify $\theta(x)$, we perform congruent diagonalization on matrix A. Let $Q=\begin{pmatrix}\frac{2}{3}&0&1\\-\frac{1}{3}&\frac{1}{\sqrt{3}}&1\\-\frac{1}{3}&-\frac{1}{\sqrt{3}}&1\end{pmatrix}$, then $\frac{1}{2}Q^TAQ=\Lambda$, where $\Lambda=\begin{pmatrix}1&0&0\\0&1&0\\0&0&0\end{pmatrix}$. For any arbitrary x, there always exists a unique s such that $x=Qs$. Substitute this into the expressions of $\theta(x)$:

We obtain $\theta(x)=\frac{(1,0,0)s}{\sqrt{s^T\Lambda s}}=\frac{s_1}{\sqrt{s_1^2+s_2^2}}$, where s_1 and s_2 are not 0 simultaneously. Let $x^e=Qt$, and substitute it into the first equation of the system (15):

We obtain $\frac{s_1}{\sqrt{s_1^2+s_2^2}}=\frac{t_1}{\sqrt{t_1^2+t_2^2}}$, After simplification, we get $(t_1,t_2,t_3)^T=(ms_1,\pm ms_2,n)^T$, where m is a positive real number and n is any real number.

Substitute $x^e=Qt$ and $s=Q^{-1}x$ into the above equation, we have $x^e=\frac{m}{3}Ax+n(1,1,1)^T$ or $x^e=\frac{m}{3}Ax+n(1,1,1)^T$, where $E(2,3)=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}$.

 doi:10.20944/preprints202510.0998.v1

The following is the discussion by cases:

1. First item; $x^e = \frac{m}{3}Ax + n(1,1,1)^T$

In this case, the equation $\begin{cases} R^e - G^e = m(R-G) \\ R^e - B^e = m(R-B) \\ G^e - B^e = m(G-B) \end{cases}$ hold, thus the magnitude relationship of R^e, G^e, B^e is the same as that of R, G, B. Let $\min(x) = x_i$,Then we have: $\min(x^e) = x_i^e$, where $(x_1^e, x_2^e, x_3^e) = (R^e, G^e, B^e)$
$(x_1, x_2, x_3) = (R, G, B)$.

Denote $M(i)$ as the i - th row of any matrix M. Then $\min(x) = E(i)x$, (where E is the 3×3 identity matrix), so $S(x) = 1 - \frac{3E(i)x}{(1,1,1)x}$.

Substitute the above formula into the second equation of the system (15), we get $\frac{E(i)x}{(1,1,1)x} = \frac{E(i)x^e}{(1,1,1)x^e}$. Substitute the expression of x^e into this equation, and considering that $(1,1,1)x^e = 3n$, the equation simplifies to $n[3E(i) - (1,1,1)]x = \frac{m}{3}A(i)x(1,1,1)x$. Further, considering that $3E(i) - (1,1,1) = A(i)$, the equation finally simplifies to $n = \frac{m}{3}(1,1,1)x$.

Substitute it into the expression of x^e, we obtain $x^e = mx, m > 0$.

Substitute it back into the original system (15) for verification, and it holds.

2. Second item; $x^e = \frac{m}{3}AE(2,3)x + n(1,1,1)^T$

Note that $x_2^e - x_3^e = m(x_3 - x_2)$, i.e., $G^e - B^e = m(B-G)$, so the magnitude relationship of G, B is opposite to that of G^e, B^e.

In this case, $H(x) = 360^\circ - H(x^e)$, and the original system (15) does not hold, so this case is discarded.

To sum up, the system (15) holds if and only if $x^e = mx, m > 0$, which means the two pixels have the same H and S.

Now, normalize the pixel values and let $y = (Y, Cb - 0.5, Cr - 0.5)^T$ and $y^e = (Y^e, Cb^e - 0.5, Cr^e - 0.5)^T$ .From Equation (9), we have $y = Bx$, where $B = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.168736 & -0.331264 & 0.5 \\ 0.5 & -0.418688 & -0.081312 \end{pmatrix}$.

Note that matrix B is invertible, so if $x^e = mx, m > 0$ is to hold, then $y^e = my, m > 0$ must hold. □

References

1. Shangchen Zhou,Chongyi Li & Chen Change Loy. LEDNet: Joint Low-Light Enhancement and Deblurring in the Dark[C]//Computer Vision – ECCV 2022. 2022.
2. Hao Zhang,Jiayi Ma. SDNet: A Versatile Squeeze-and-Decomposition Network for Real-Time Image Fusion[J]. International Journal of Computer Vision,2021,Vol.129(10): 2761-2785.

 doi:10.20944/preprints202510.0998.v1

3. Han Xu, Jiayi Ma, Junjun Jiang,et al. U2Fusion: A Unified Unsupervised Image Fusion Network[J]. IEEE transactions on pattern analysis and machine intelligence,2022,Vol.44(1): 502-518.
4. Hui Li,Xiao-Jun Wu,Josef Kittler. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. Information Fusion,2021,Vol.73: 72-86.
5. Xu, Han,Zhang,et al. Classification saliency-based rule for visible and infrared image fusion.[J]. IEEE Trans. Comput. Imaging,2021,Vol.7: 824-836.
6. Jiayi Ma,Wei Yu,Pengwei Liang,et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information Fusion,2019,Vol.48: 11-26.
7. Jiayi Ma[1],Hao Zhang[1],Zhenfeng Shao[2],et al. GANMcC: A Generative Adversarial Network With Multiclassification Constraints for Infrared and Visible Image Fusion[J]. Instrumentation and Measurement, IEEE Transactions on,2021,Vol.70: 1-14.
8. Linfeng Tang1,Xinyu Xiang2,Hao Zhang3,et al. DIVFusion: Darkness-free infrared and visible image fusion[J]. Information Fusion,2023,Vol.91: 477-493.
9. Jun Chen,Liling Yang,Wei Liu,et al. LENFusion: A Joint Low-Light Enhancement and Fusion Network for Nighttime Infrared and Visible Image Fusion[J]. IEEE Transactions on Instrumentation and Measurement,2024,Vol.73: 1-15.
10. Linfeng Tang1,Jiteng Yuan2,Hao Zhang3,et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware[J]. Information Fusion,2022,Vol.83: 79-92.
11. Cheng, MuhangCAa,Huang,et al. LEFuse: Joint low-light enhancement and image fusion for nighttime infrared and visible images[J]. Neurocomputing,2025,Vol.626: 129592.
12. Jia, Xinyu,Zhu,et al. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision[C]//18th IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021. 2021.
13. Liu, Jinyuan , Fan,et al. Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.
14. MA J Y, TANG L F, XU M L, et al. STDFusionNet: an infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.
15. Liu Y, Chen X, Ward R K, et al. Image fusion with convolutional sparse representation[J]. IEEE signal processing letters, 2016, 23(12): 1882-1886.
16. LI H,WU X J,KITTLER J. Infrared and visible image fusion using a deep learning framework[C]// 2018 24th International Conference on Pattern Recognition(ICPR).2018:2705-2710.
17. LI H,WU X J,DURRANI T S. Infrared and visible image fusion with ResNet and zero-phase component analysis[J].Infrared Physics & Technology,2019,102:103039.
18. LI H, WU X J. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.
19. XU H, GONG M Q, TIAN X, et al. CUFD: an encoderdecoder network for visible and infrared image fusion based on common and unique feature decomposition[J]. Computer Vision and Image Understanding, 2022, 218: 103407.
20. XU H, LIANG P W, YU W, et al. Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, Aug 10-16, 2019: 3954-3960.