第XX卷第XX期
XXXX年XX月

红 外 与 毫 米 波 学 报
J. Infrared Millim. Waves

Vol. XX, No. XX
XX,XXXX

文章编号：1001-9014(XXXX)XX-0001-13　　DOI：10. 11972/j. issn. 1001-9014. XXXX. XX. 001

BDMFuse: Multi-Scale Network Fusion for Infrared and Visible Images Based on base and Detail Features

Si Haiping[1], Zhao Wenrui[1], Li Tingting[1], Li Feitao[1], Bacao Fernando[2], Sun Changxia[1], Li Yanling[1*]

(1. College of Information and Management Science, Henan Agricultural University, Zhengzhou, Henan, 450008, China;
2. NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa)

Abstract: The result of infrared and visible image fusion should highlight the significant targets of the infrared image while preserving the visible light texture details. In order to satisfy the above requirements, this paper proposes an automated encoder-based infrared and visible image fusion method. The encoder constructs both a base encoder and a detail encoder according to the optimization objective. The base encoder extracts low-frequency information from the image, while the detail encoder captures high-frequency information. Since this extraction method may miss some information, we introduce a compensation encoder to supplement the missing information. Additionally, we introduce multi-scale decomposition for the encoder to extract image features more comprehensively. The image features obtained by the encoders are then fed into the decoder. The decoder first adds the low-frequency, high-frequency and compensatory information to obtain multi-scale features. An attention map is derived from these multi-scale features and multiplied with the fused image at the corresponding scale. The Fusion module is introduced in the multi-scale fusion process to achieve image reconstruction. The network proposed in this paper demonstrates its effectiveness on the TNO, RoadScene, and LLVIP datasets. Experiments show that our network can better perceive changes in light, effectively extract image detail information, and produce fused images that are more aligned with human visual perception.

Key words: Infrared image, Visible image, encoder-decoder, multi-scale features, attention strategy
PACS:

BDMFuse:基于红外与可见光图像基础特征和细节特征的多尺度融合

司海平[1]，赵文汭[1]，李婷婷[1]，李飞涛[1]，Fernando Bacao[2]，孙昌霞[1]，李艳玲[1*]

(1. 河南农业大学 信息与管理科学学院,郑州,河南,450008;
2. NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa)

摘要:红外与可见光图像融合结果应该突出红外图像显著目标,同时保留可见光纹理细节。为了满足上述要求,本文提出一种基于自编码器的红外与可见光图像融合方法。编码器根据优化目标构建基础编码器和细节编码器。基础编码器用于提取图像的低频信息,细节编码器用于提取图像的高频信息。这种提取方式会造成部分信息未被捕捉,我们提出补偿编码器补充信息。同时,我们引入多尺度分解来更加全面的提取图像特征。随后,将编码器获取的图像特征送入解码器。解码器先将高频、低频和补充信息相加获取多尺度特征。我们从多尺度特征中获取注意力图,与对应尺度的融合图像相乘。多尺度融合中引入Fusion模块,实现图像重建。本文提出的网络在TNO、RoadScene和LLVIP数据集上证明网络的有效性。实验表明我们的网络能够更好的感知光线变化,较好的提取图像细节信息,融合后的图像更符合人的视觉感知。

关　键　词:红外图像;可见光图像;编码器-解码器;多尺度特征;注意力策略

Foundation items: Supported by the Henan Province Key Research and Development Project (231111211300), The central government of Henan Province guides local science and technology development funds(Z20231811005), Henan Province Key Research and Development Project (231111110100) and Henan Provincial Outstanding Foreign Scientist Studio(GZS2024006).

Biography: Haiping Si(1987-), male, professor. Research interests include image fusion. E-mail: haiping@henau. edu. cn.
[*] **Corresponding author**: E-mail: lyl_lingling@163. com.

1 Introduction

Infrared and visible image fusion is a vital task in image processing [1]. Infrared images, captured through thermal radiation, are less affected by external factors but often suffer from a significant loss of texture details and structural information [2]. In contrast, visible images contain rich texture information but are more sensitive to environmental changes [3]. Therefore, combining infrared and visible images in a bimodal manner can generate high-quality images [4].

To maximize the retention of useful information in fused images, various methods have been developed. Traditional image fusion techniques operate in either the spatial domain or the transform domain. Spatial domain methods directly process image pixels, whereas transform domain methods handle the frequency representation of images using transformations such as the Fourier transform or wavelet transform [5-6]. While these traditional methods have produced satisfactory fusion results, they largely depend on heuristic fusion rules. Consequently, traditional image fusion techniques often fall short of meeting the increasingly stringent fusion requirements.

In recent years, the feature extraction capabilities of deep learning have garnered significant attention from researchers. Studies have demonstrated that deep learning methods can substantially enhance the quality of fused image [7]. Currently, deep learning-based image fusion methods can be broadly categorized into three types: autoencoder (AE) -based methods, convolutional neural network (CNN)-based methods, and generative adversarial network (GAN) -based methods. AE-based methods focus on encoding data to create a low-dimensional representation, which is then decoded to reconstruct the original data [8]. CNN-based methods extract image features through local connections and shared weights [9]. GAN-based methods employ adversarial training between a generator and a discriminator to produce high-quality images [10].

Although methods for infrared and visible image fusion are relatively mature, there are still unresolved issues. Increasing the number of network layers enhances the network's expressive ability, but it also exacerbates the loss of image detail. This, to some extent, limits the further improvement of fusion image quality. Our proposed network effectively addresses this problem, with its main contributions being as follows:

(1) The encoder network constructs base and detail encoders to extract low and high frequency information based on the optimisation objective, and constructs compensation encoders to supplement the information.

(2) The decoder network first fuses the different scales of low-frequency, high-frequency, and compensatory information to obtain multi-scale features. These multi-scale features are then multiplied by the acquired attention map, and the Fusion module is introduced to perform multi-scale fusion for image reconstruction.

(3) Our proposed fusion method demonstrates superior fusion results in terms of both visual assessment and objective evaluation compared to nine other fusion algorithms on three public datasets.

2 Related work

2.1 Methods Based on Autoencoder

An autoencoder is an unsupervised learning model commonly used for tasks such as data compression and feature extraction. The encoder maps input data into a low-dimensional feature space, while the decoder reconstructs the original data from this reduced space. Through this process, the autoencoder retains essential information and effectively captures significant feature representations [11]. As a result, autoencoder have been widely applied in areas such as target detection [12], target segmentation [13], and early warning systems [14].

In image processing, the feature extraction capabilities of self-encoders are extensively utilized in image fusion tasks. Image fusion aims to combine information from multiple image sources to generate richer and more informative results. Autoencoder is usually divided into three parts in this field: the encoder, the decoder, and the fusion layer. The encoder automatically learns the feature information of the source images without the need for manually designed features. The decoder maps the low-dimensional features extracted by the encoder back to the original space, adaptively reconstructing the image. Typically, image fusion methods based on autoencoders train the encoder and decoder during the training phase. In the testing phase, a fusion layer, such as an addition or maximum strategy, is introduced to merge the image features of infrared and visible.

Li et al. [15] proposed a deep learning architecture that integrates convolutional layers, fusion layers, and dense blocks, where the output of each layer is connected to the outputs of other layers. Recognizing that the features extracted by a single branch might lack comprehensiveness, Li et al. [16] introduced a method incorporating a multi-level residual encoder module and a decoder module with hybrid transmission. This design features a multi-level residual encoder module with two independent branches for extracting image features. Zhao et al. [17] proposed a dual-branch structure for multimodal feature decomposition and image fusion. Tang et al. [18] proposed a darkness-free infrared and visible image fusion method, which fully considers the intrinsic relationship between low-light image enhancement and image fusion, achieving effective coupling and information complementarity. Additionally, Tang et al. [19] were the first to consider the gap between high-level vision tasks and image fusion, proposing a semantic-aware image fusion framework.

2.2 Multi-scale transform

Multiscale transform is a technique for processing and analyzing information by decomposing an image or signal into different scales or frequency components[20]. It can effectively capture local details and global features in images and is widely used in tasks such as image fusion, texture analysis and feature extraction. The commonly used multiscale transforms mainly contain methods such as pyramid transformations [21], wavelet transformations [22], and non-subsampled multi-scale multi-directional geometric transformations [23].

With the advancement of deep learning, researchers and scholars have integrated multiscale transforms into deep learning-based networks. Lin et al [24] proposed cross-scale fusion module for interactive fusion of features between encoder and decoder. Jian et al. [25] introduced a multiscale encoder to extract featured image features and constructed a symmetric encoder-decoder with residual blocks (SEDRFuse) network for fusing infrared and visible images in night vision applications. Wang et al. [26] introduced a novel and efficient fusion network based on dense Res2net and dual nonlocal attention models. They integrated Res2net and dense connectivity into an encoder network, enabling the utilization of multiple available receptive fields to extract multiscale features. This approach aims to retain as much effective information as possible for the fusion task.

3 proposed fusion method

This section provides a detailed description of the proposed architecture for infrared and visible image fusion, encompassing the encoder, decoder, and training specifics.

3.1 Network architecture

Our proposed Infrared and Visible Image Fusion Architecture (BDMFuse) is an end-to-end network architecture. The encoder employs three branches: the base encoder, detail encoder, and compensation encoder. The base encoder extracts low-frequency information from the image, the detail encoder captures high-frequency information, and the compensation encoder gathers information not captured by the other encoders. Each encoder operates at three different scales, and their outputs are fed into the decoder. Multi-scale features are first generated by summing the low-frequency, high-frequency, and compensatory information from different scales. These multi-scale features are then fused to achieve image reconstruction.

3.2 Encoder

Inspired by the literature [27], we construct the base encoder, detail encoder, and compensation encoder with the idea of solving for the optimal solution. The base encoder extracts low-frequency features by minimizing Eq. (1) to obtain the optimal solution.

$$B_n = B_I - \gamma_{B1} * B_m, \qquad (1)$$

Where B_I is the image acquired after applying a blur filter to the input image and γ_{B1} is the tuning hyperparameter. B_n (n∈{1,2,3}) is the decomposed base image. B_m (m∈{1,2,3}) is the low-frequency features of different scales obtained after convolution operations of different depths, which are represented as shown below:

$$B_m = L_n - \gamma_{B1} * E_n, \qquad (2)$$

where L_n (n∈{1, 2, 3}) denotes features of different scales obtained after convolution operations of different depths, E_n (n∈{1,2,3}) denotes high-frequency features of different scales extracted from the E_n feature map, and γ_{B2} is the tuning hyperparameter.

The detail encoder minimizes Eq. (3) to obtain the optimal solution.

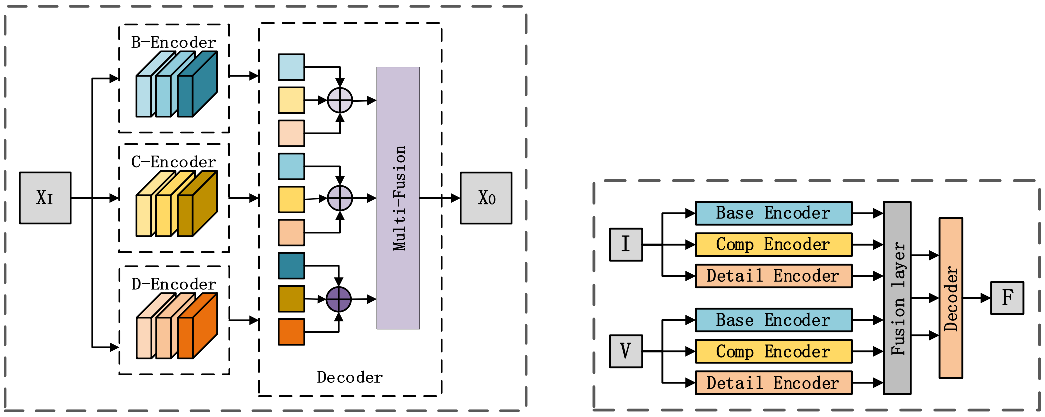
$$D_n = D_I - \gamma_{D1} * D_m, \qquad (3)$$

Where D_I is the image obtained after applying Laplace filter to the input image and D_m (m∈{1,2,3}) denotes the high frequency features at different scales obtained from the convolution operation that has been performed at different depths. γ_{D1} is the tuning hyperparameter. D_n (n∈{1,2,3}) is the decomposed detail image

The compensation encoder also adopts the method of finding the optimal value to obtain the feature information, which is similar to the detail encoder and will not be repeated. We obtain the compensating filter by Eq. (4).

$$C_I = 1 - B_I - D_I. \qquad (4)$$

The base encoder network is depicted in Fig. 2. We



(a) Overall structure of the training phase.

(b) Overall structure of the testing phase.

Fig.1 The overall network of BDMFuse.
图 1 BDMFuse的整体网络架构

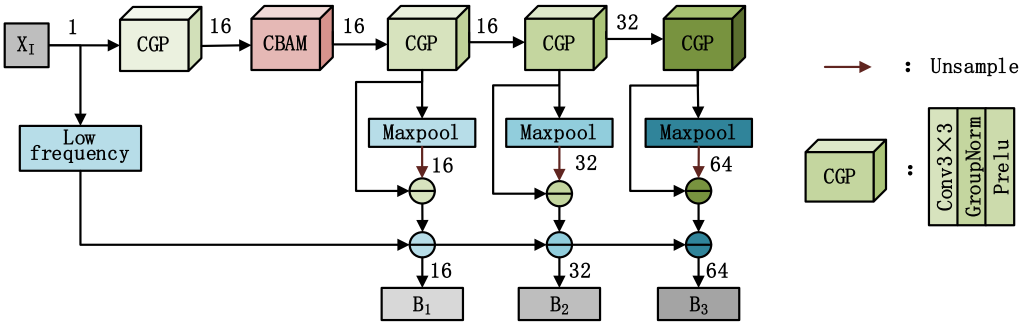


Fig. 2 Base encoder schematic.
图2 基础编码器架构

adopt two processing methods to extract the low-frequency information of the input image. One is to take low pass filter and the other is to take convolution operation. Firstly, feature coarse extraction is performed, followed by the addition of the CBAM attention mechanism [28] to emphasize effective information. To capture more comprehensive information, we introduce multi-scale feature extraction. Maximum pooling and up-sampling operations are applied to features of different scales to extract high-frequency information, thereby removing high-frequency components from the features. The pooling operation reduces the feature size by half, and to maintain the image size, bilinear interpolation-based up-sampling is adopted.

The detail encoder network is illustrated in Fig. 3, and its architecture bears resemblance to the base encoder. However, the detail encoder employs two methods to obtain the high-frequency information of the image. One approach is to utilize a high-pass filter, while the other involves convolutional operations. Maximum pooling operation is taken directly on image features of different scales in convolution operation to obtain the high frequency information of the image. In this process, the CGP convolution blocks in the network all employ 3×3 convolution kernels with a stride of 1. Transitioning through the CGP convolution block alters only the number of channels and does not affect the image size.

The compensation encoder is shown in Fig. 4. This encoder is constructed to complement the image feature information, so only the basic convolutional processing of the rest of the encoder is retained without any other additional operations.

3.3 Decoder

3.3.1 Decoder Network

Initially, the corresponding low-frequency information and high-frequency information of different scales are fused. As shown in Fig. 5(a), F_b^n(n∈{1,2,3}) refers to the multi-scale low-frequency features, F_d^n(n∈{1,2,3}) denotes the multi-scale high-frequency features, and F_c^n(n∈{1,2,3}) represents the multi-scale compensation features. Since the compensating features may increase image artifacts, we add hyperparameters β to modulate the multi-scale compensating features. The size of F_d^n differs from the rest of the image sizes, so up-sam-

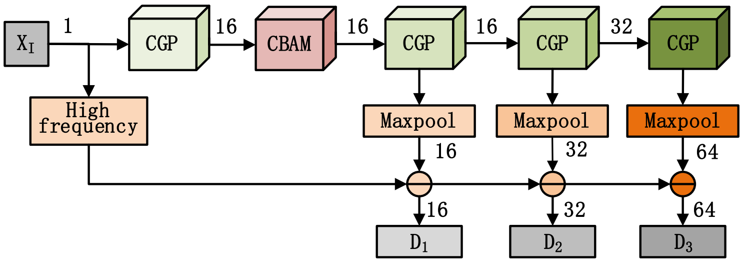


Fig. 3 Detail encoder schematic.
图3 细节编码器架构

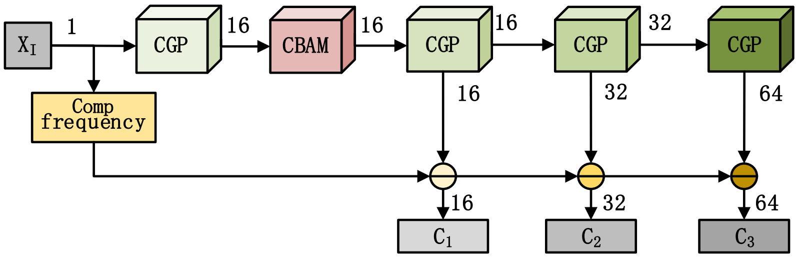


Fig. 4 Compensation encoder schematic.
图4 补偿编码器架构

pling is achieved through an inverse convolution operation for F_d^n. Finally, F_b^n, F_d^n, and F_c^n are summed to obtain the multi-scale feature F_n(n∈{1,2,3}).

In order to be able to make the image learn more multi-scale information, we obtain the attention map of the corresponding scale from F_n and multiply it with the fused image. Then multi-scale fusion is performed, as shown in Fig. 5(b). F_3 is spliced with F_2 after the CGS module is aligned with the number of F_2 channels, and then sent to Fusion to achieve fusion. After F_3 and F_2 are fused, they are passed through CGS to maintain the same number of channels as F_1. Then, they are concatenated with F_1 and sent into the Fusion module. The Fusion module [29] is shown in Fig. 6. We introduce hyperparameters to the Fusion module to enhance its applicability to our proposed network. Finally, image reconstruction is achieved by two CGS modules. The CGS convolutional blocks in the network adopt 3×3 convolutional kernels with a step size of 1. The CGS convolutional blocks only change the number of channels and do not change the image size.

3.3.2 Attention strategy

Our computation of F_n weights draws inspiration from ASFF (Adaptively Spatial Feature Fusion) [30]. F_n has the same size but different numbers of channels. We adjust different number of channels for F_n to obtain three attention maps to extract the weight values of the corresponding scales. The formula is shown below:

$$\omega_1^n,\omega_2^n,\omega_3^n = softmax\left(F_{1\rightarrow n},F_{2\rightarrow n},F_{3\rightarrow n}\right),(\mathrm{n,i} \in \{1,2,3\}), \tag{5}$$

Where, ω_1^n, ω_2^n, ω_3^n denote the individual weights of F_n from the *n*th layer, and $F_{1\rightarrow n}$, $F_{2\rightarrow n}$, $F_{3\rightarrow n}$ denote the number of channels of all the layers adjusted to the number of channels of the *n*th layer by the CGP convolution block. The softmax function is shown below:

$$softmax\left(F_{1\rightarrow n},F_{2\rightarrow n},F_{3\rightarrow n}\right) = \frac{e^{F_{i\rightarrow n}}}{e^{F_{1\rightarrow n}} + e^{F_{2\rightarrow n}} + e^{F_{3\rightarrow n}}},(\mathrm{n,i} \in \{1,2,3\}). \tag{6}$$

The weights corresponding to F_n are multiplied with F_n to optimise the F_n features as follows:

$$F_n = F_n{*}\omega_n^n, \tag{7}$$

where ω_n^n is the corresponding weight of $\mathrm{F_n}$ in the nth layer of the attention graph.

Taking Attention-3 as an example, Fig. 7 illustrates how to obtain the weights of F_3. We input F_n into the CGP convolution block, adjust the number of channels of F_1 and F_2 to match that of F_3, and obtain the feature maps. Subsequently, the three feature maps are concatenated, and the weight value of each feature map is calculated using softmax. Finally, the corresponding weight of F_3 is extracted and multiplied with F_3 to optimize the F_3 features. In this process, the CGP convolution blocks all employ 1×1 convolution kernels with a stride of 1.

3.4 Training Strategy

Our training strategy is similar to DenseFuse. In the training phase, the fusion network is discarded. Through training, we expect the encoder to extract multi-scale depth features and the decoder to reconstruct the image based on these features. This training strategy can leave more choice space for the fusion layer.

In the training phase, the loss function L_{total} is defined as follows:

$$L_{total} = L_p + \lambda L_s + \lambda/2(L_{s2} + L_{s1}), \tag{8}$$

Where L_p and L_s denote the pixel loss and structural similarity loss between the input image and the output image respectively. λ is a measure between L_p, L_s, L_{s1} and L_{s2}.

L_p is calculated by Eq. (9):

$$L_p = \|O - I\|^2, \tag{9}$$

where O and I denote the output and input images, respectively. L_p calculates the square of the difference between the output image and the input image, aiming to ensure that the reconstructed image is similar to the source image at the pixel level.

L_s is obtained from Eq. (10):

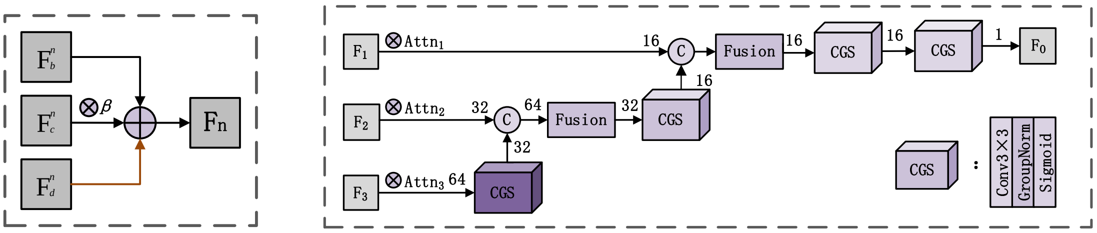
$$L_s = 1 - SSIM(O,I), \tag{10}$$

Where SSIM(-) represents the structural similarity measure. When the value of SSIM(-) is larger, it indicates a higher structural similarity between the output image O and the input image I.

The Fusion module is introduced in the decoder network to retain more feature information. However, it tends to ignore the structural information of the image to some extent. To address this, L_{s2} and L_{s1} losses are introduced.

L_{s2} is obtained from Eq. (11):

$$L_{s2} = 1 - SSIM(f_{3+2},F(f3,f2)), \tag{11}$$



(a) BDC-Fusion

(b) Multi-Fusion

Fig. 5 Decoder schematic.
图5 解码器网络

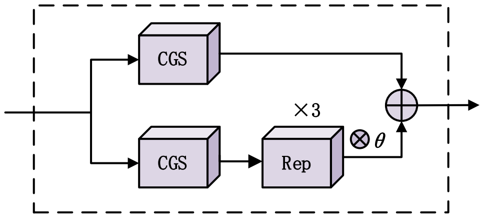


Fig. 6 Fusion Module.
图6 Fusion模块

Where f_{3+2} is the result of summing the multiscale features F_2 and F_3, while $F(f3,f2)$ is the fusion result obtained after the Fusion module.

L_{s1} is obtained from Eq. (12):

$$L_{s1} = 1 - SSIM(f_{2+1}, F(f2,f1)), \qquad (12)$$

Where f_{2+1} is the result of fusion of multiscale features F_2 and F_3 and then summed with F_1, and $F(f2,f1)$ is the fusion result obtained after the Fusion module.

The goal of the training phase is to train an autoencoder network capable of effectively extracting image features and reconstructing images. We employ 1600 images from FLIR as input images for network training. The training samples are randomly cropped to 224 × 224. we set the parameter λ to 10 to train the network.

4 Experimental results and analysis

In this section, we will validate our proposed fusion method through a series of experiments. The network is executed on an NVIDIA GeForce RTX 2080Ti and implemented using the PyTorch framework.

4.1 Experimental setting

In this experiment, images are selected from the TNO, RoadScene[31], and LLVIP[32]datasets, and our proposed fusion network is compared with nine typical fusion networks (DenseFuse, RFN-Nest, FusionGAN [33], U2Fusion [34], CSF [35], SEDR , SwinFusion [36], SeAFusion, CDDFuse) to evaluate its performance. We select seven commonly used image quality evaluation metrics for objective quality evaluation of fused images, which are entropy (EN) [37]; mutual information (MI) [38]; standard deviation (SD) [39]; average gradient (VG) [40], visual information fidelity (VIF) [41]; the sum of the correlations of differences (SCD) [42]; multiscale structural similarity measure (MS_SSIM) [43]. Among them, EN is an objective evaluation metric that measures how much information the image contains; MI is used to measure the amount of information transferred from the source image to the fused image; SD reflects the distribution and contrast of the fused image; AG is a metric used to evaluate the sharpness and detail of an image; VIF measures the information fidelity of the fused image; SCD is a measure of the differences between the fused image and the original image; and MS_SSIM is a similarity-based evaluation metric.

4.2 Ablation experiment

The proposed network incorporates compensated encoders into the encoder network and introduces an attention strategy along with a Fusion module for multi-scale fusion in the decoder network. To validate the efficacy of these strategies, the article includes ablation experiments. Additionally, to further assess the benefits of the attention strategies, the article compares two attention mechanisms SE and ECA with the strategies we have adopted. As illustrated in Fig. 8, our proposed network demonstrates superior visual fusion effects in the fused image. Specifically, it effectively highlights the presence of the villain in the infrared image while preserving the dendritic texture features in the visible image.

To further demonstrate the superiority of our method, we conducted a quantitative analysis using test results from 40 pairs of infrared and visible images selected from the TNO dataset. The optimal values are highlighted in bold black. As shown in Table 1, our proposed network achieves optimal values for the EN, MI, SD, and VIF metrics. These results provide additional evidence of the effectiveness of our adopted approach.

4.3 Analysis of experimental results

4.3.1 Subjective analysis

Subjective qualitative comparisons of the nine existing fusion methods with our fusion method are presented

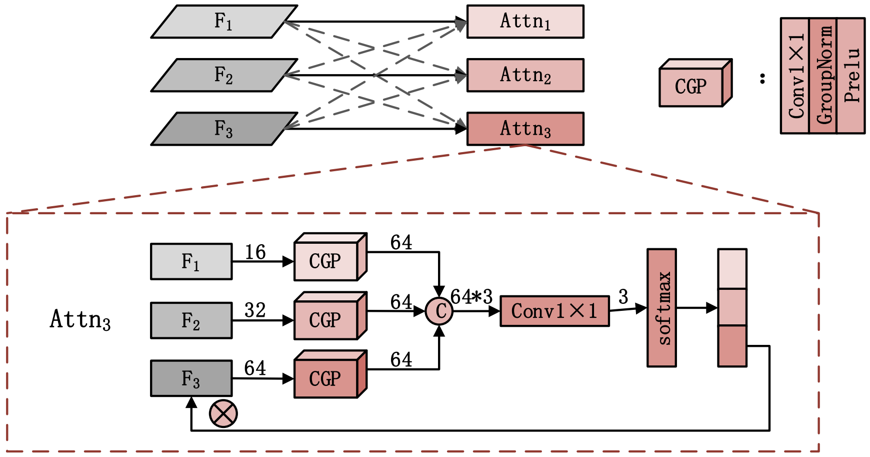
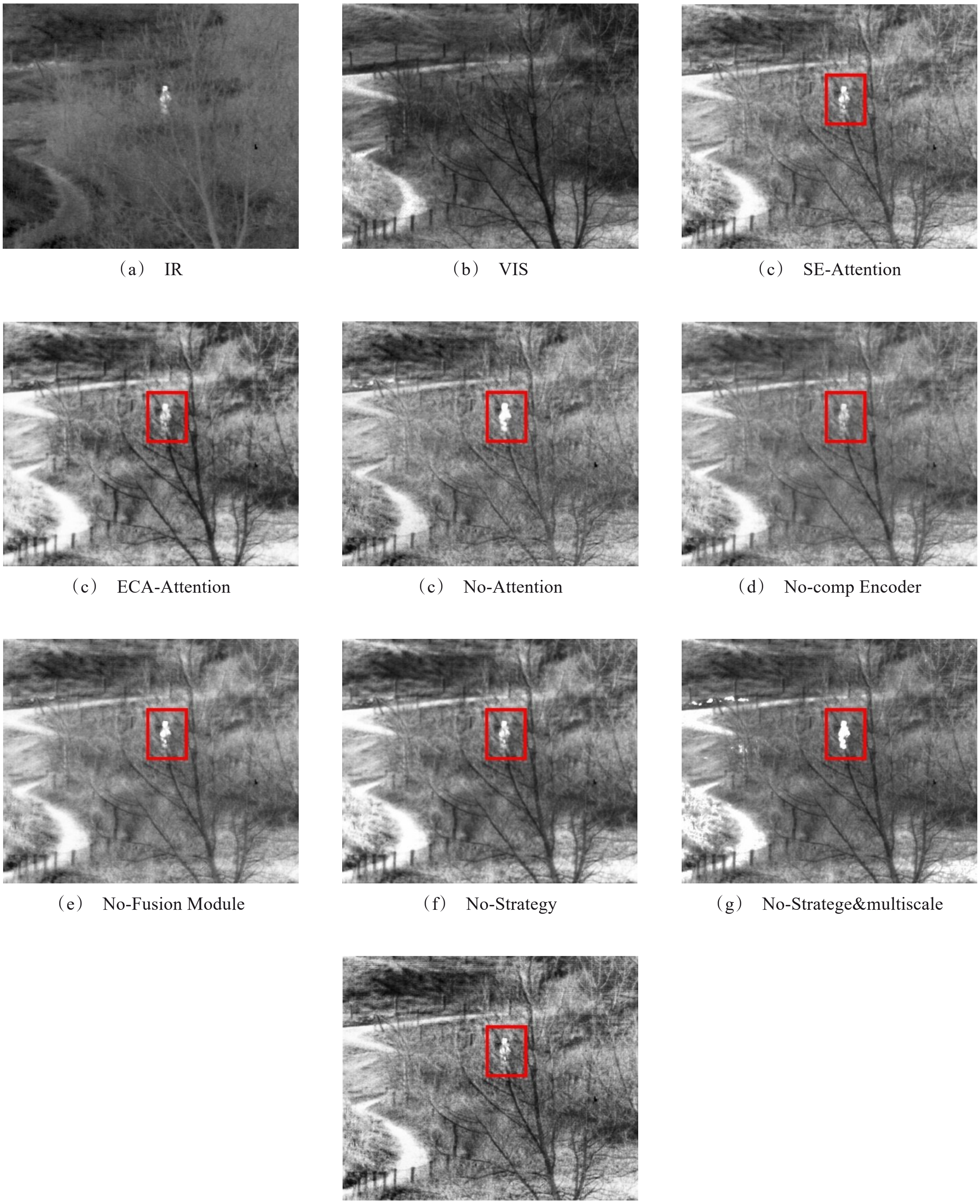


Fig. 7 Attention computation strategy
图7 注意力计算策略


（a） IR

（b） VIS

（c） SE-Attention

（c） ECA-Attention

（c） No-Attention

（d） No-comp Encoder

（e） No-Fusion Module

（f） No-Strategy

（g） No-Stratege&multiscale

（h） ours

Fig. 8 Fusion image of ablation experiments
图8 消融实验的融合结果

in Figures 9~11. As depicted in the figures, all fusion methods effectively fuse infrared and visible images, but notable differences in visual quality are observed. FusionGAN retains more information from the infrared image during fusion, which can result in blurred image details. The remaining fusion methods retain more detailed

Table 1 Average quality evaluation metrics for ablation experiments
表1 消融实验的平均质量评估指标

	EN	MI	SD	AG	VIF	SCD	MS_SSIM
SE-Attention	6. 8501	13. 7001	34. 7315	3. 7415	0. 9612	1. 8192	**0. 9416**
ECA-Attention	6. 9504	13. 9008	40. 6716	3. 5503	0. 9548	1. 6514	0. 9148
No-Attention	7. 1774	14. 3547	40. 2576	**4. 2096**	1. 3605	1. 8801	0. 9306
No-comp Encoder	6. 8728	13. 7456	38. 0053	2. 9682	0. 9219	1. 8060	0. 9403
No-Fusion Module	7. 2129	14. 4259	40. 8263	3. 7076	1. 2727	1. 8870	0. 9277
No-Strategy	7. 0291	14. 0581	37. 9805	3. 3666	1. 0851	**1. 8935**	0. 9511
No-Strategę&multiscale	6. 5910	13. 1819	35. 0411	3. 2785	0. 8063	1. 6111	0. 9216
ours	**7. 2525**	**14. 5050**	**43. 6183**	4. 1627	**1. 4128**	1. 8715	0. 9207

information from the visible image, especially SwinFusion, SeAFusion and CDDFuse. This method produces fusion results that align more closely with subjective human visual perception but experience some loss of saliency information. For example, in Fig. 9, the cloud information present in the infrared image is missing. Our proposed network retains more light and shadow information, as illustrated in Figure 10, indicating its enhanced ability to perceive light. To further validate this, we conducted experiments on the LLVIP dataset. LLVIP is a dataset for low-light vision paired with visible and infrared images. As demonstrated in Figure 11, our fused images reveal the basic outline of the manhole cover even under low-light conditions.

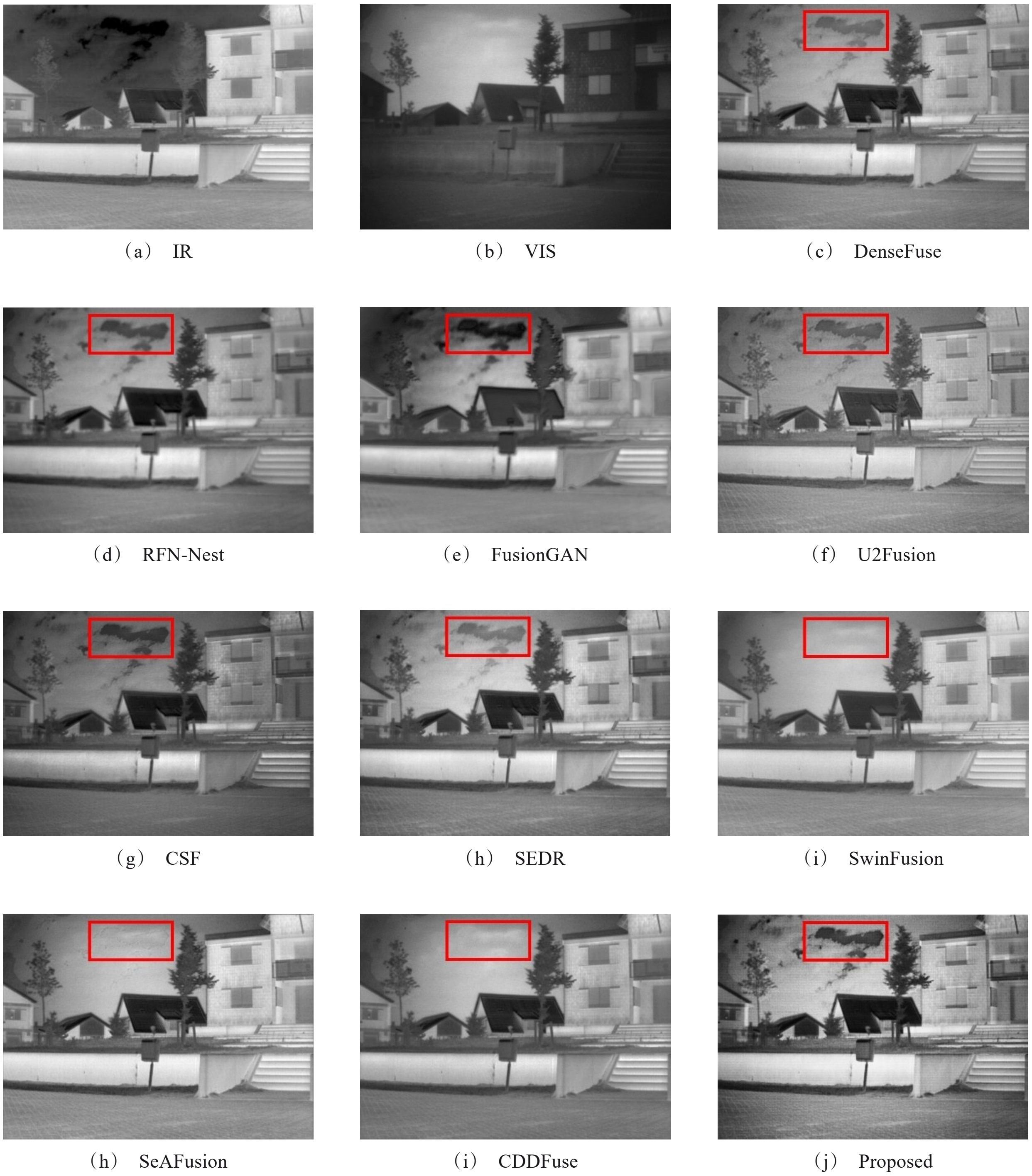
4. 3. 2 Quantitative analysis

In addition to subjective qualitative analysis, we employ objective quantitative analysis to measure the performance of the proposed method. The quantitative evaluation results for the test sets of 40 pairs from TNO, 100 pairs from RoadScene, and 100 pairs from LLVIP are shown in Tables 2-4. The best value is indicated in bold black text, and the second-best value is underlined. Using the experimental results of the LLVIP dataset as an example, we achieve the highest values on five indicators: EN, MI, SD, VIF and SCD. This demonstrates our network's capability to effectively extract image details. The results from the other two datasets show that while our proposed fusion method may fall slightly short of the optimal values in some evaluation metrics, the outcomes are relatively balanced across all metrics and generally superior to other fusion methods. Additionally, the experimental results from all three datasets indicate that our method has good feasibility and generalizability.

5 Conclusion

This paper designs a multi-scale transformation fusion method based on fundamental and information extraction. We construct the base encoder and detail encoder based on optimization problems to extract low-frequency and high-frequency information from images. Additionally, a compensation encoder is proposed to supplement the information. To obtain more image feature information, we introduce multi-scale extraction of image information. The decoder first adds the low-frequency, high-frequency, and compensatory information to obtain fused images at different scales. Attention maps are then derived from these fused images, and the corresponding weights are multiplied with the fused images at different scales. Finally, the Fusion module is introduced for multi-scale fusion to achieve image reconstruction. During the training phase, the three encoders and one decoder are trained according to the loss function to ensure image reconstruction capability. In the testing phase, the low-frequency, high-frequency, and compensatory information at different scales decomposed by the encoder are fed into the fusion layer to integrate the corresponding infrared and visible images. Finally, these are sent into the decoder for image reconstruction, resulting in the final fused image.

Our network undergoes comparative experiments on the TNO, RoadScene, and LLVIP datasets. The results demonstrate that both the subjective fusion effect and the overall objective evaluation of our proposed network outperform those of the comparison methods, with good generalization across datasets.



（a） IR

（b） VIS

（c） DenseFuse

（d） RFN-Nest

（e） FusionGAN

（f） U2Fusion

（g） CSF

（h） SEDR

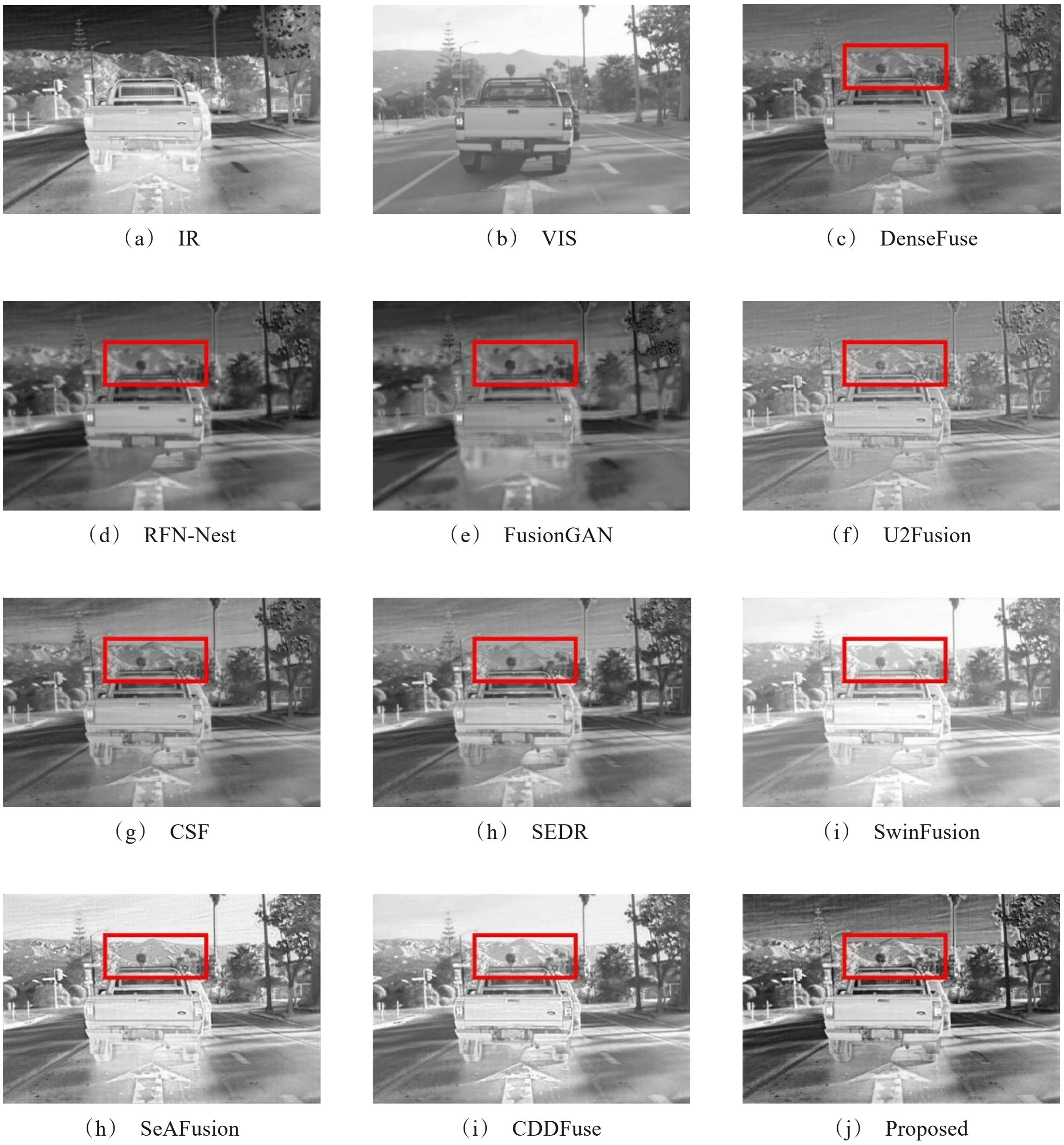
（i） SwinFusion

（h） SeAFusion

（i） CDDFuse

（j） Proposed

Fig. 9 Fusion image of TNO
图9 TNO图像的融合结果



（a） IR

（b） VIS

（c） DenseFuse

（d） RFN-Nest

（e） FusionGAN

（f） U2Fusion

（g） CSF

（h） SEDR

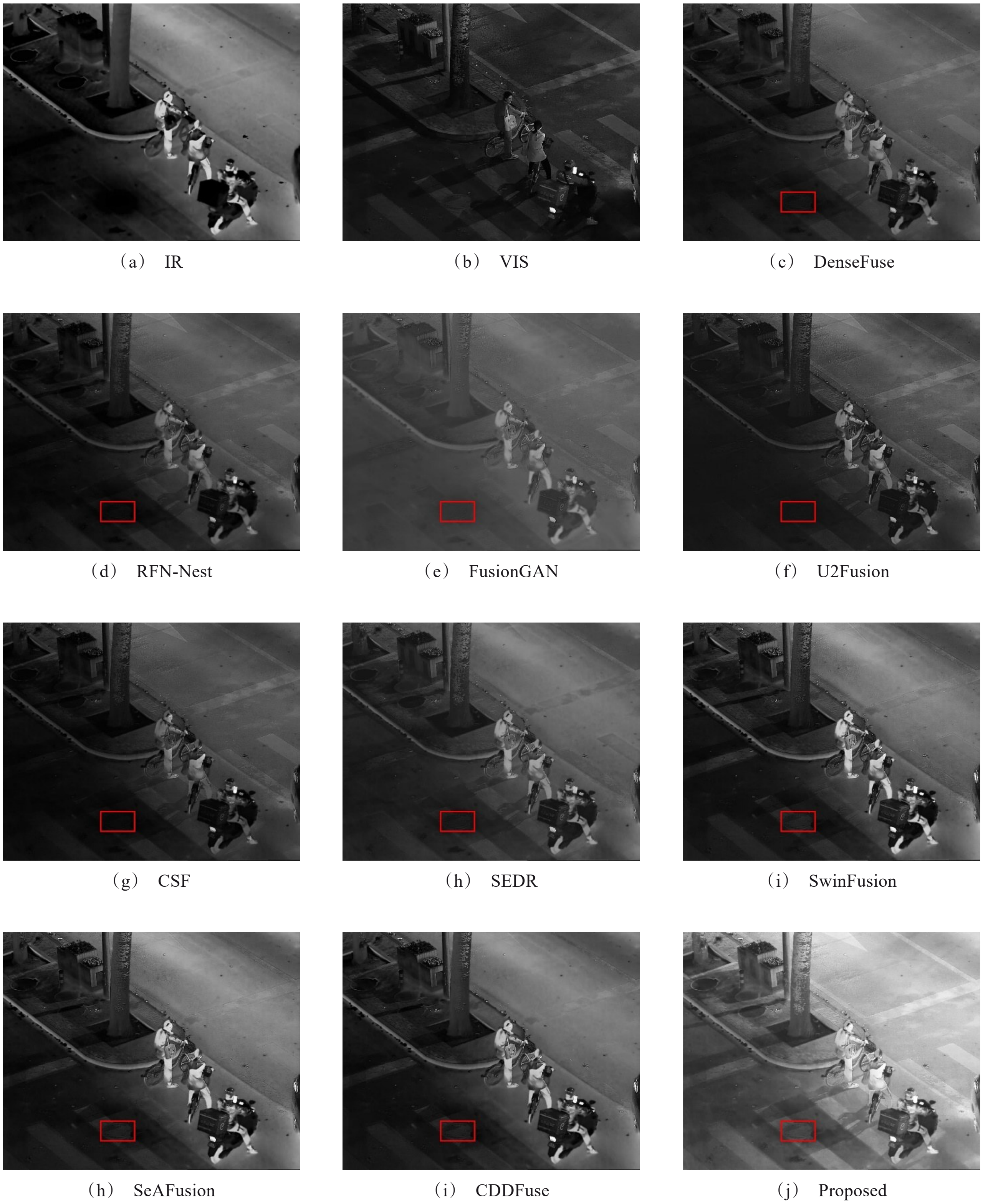
（i） SwinFusion

（h） SeAFusion

（i） CDDFuse

（j） Proposed

Fig. 10 Fusion image of RoadScene
图10 RoadScene图像的融合结果



（a） IR

（b） VIS

（c） DenseFuse

（d） RFN-Nest

（e） FusionGAN

（f） U2Fusion

（g） CSF

（h） SEDR

（i） SwinFusion

（h） SeAFusion

（i） CDDFuse

（j） Proposed

Fig. 11 Fusion image of LLVIP
图11 LLVIP图像的融合结果

Table 2 Average quality evaluation metrics for 40 pairs of TNO fused images

表2 40对TNO融合图像的平均质量评估指标

	EN	MI	SD	AG	VIF	SCD	MS_SSIM
DenseFuse	6.7933	13.5867	33.0055	3.8958	1.1141	**1.8951**	0.9379
RFN-Nest	6.9889	13.9777	35.2026	2.8619	1.1218	1.8710	0.9138
FusionGAN	6.5073	13.0147	26.4478	2.4276	0.7408	1.1339	0.7511
U2fusion	6.4745	12.9489	24.9858	3.8414	0.7371	1.6542	**0.9433**
CSF	6.9295	13.8591	34.2342	3.8565	1.2221	1.8503	0.9190
SEDR	6.8795	13.7590	39.0527	4.1491	1.5546	1.8554	0.8946
swinfusion	6.6156	13.2311	31.0339	3.4971	0.7228	1.7147	0.8992
seAFusion	7.1474	14.2949	39.9100	5.6162	1.7369	1.7323	0.8553
CDDFuse	7.0582	14.1164	39.1751	5.2135	1.3976	1.7966	0.8795
ours	**7.3544**	**14.7088**	**44.5808**	**5.9907**	**1.8274**	1.8784	0.8935

Table 3 Average quality evaluation metrics for 100 pairs of RoadScene fused images

表3 100对RoadScene融合图像的平均质量评估指标

	EN	MI	SD	AG	VIF	SCD	MS_SSIM
DenseFuse	7.2684	14.5368	43.1679	4.5979	0.6504	1.6552	0.9220
RFN-Nest	7.3492	14.6984	46.1085	3.1668	0.6091	1.6837	0.8671
FusionGan	7.0540	14.1079	39.0645	3.2246	0.4805	1.0496	0.7547
u2fusion	7.0815	14.1629	37.8472	5.2514	0.6338	1.3866	0.9148
CSF	7.4257	14.8514	47.9828	5.0241	0.7952	1.7282	0.9261
SEDR	7.4499	14.8999	49.3581	4.9322	0.8286	1.6978	0.9005
swinfusion	6.9712	13.9424	45.1634	4.3312	0.7275	1.5772	0.8470
seAFusion	7.5117	15.0234	56.1119	**6.8737**	1.0949	1.6732	0.8786
CDDFuse	7.5003	15.0007	**56.4331**	6.5083	**1.1120**	1.7105	0.8740
ours	**7.5623**	**15.1246**	53.1000	6.5506	1.0359	**1.7766**	**0.9353**

Table 4 Average quality evaluation metrics for 100 pairs of LLVIP fused images

表4 100对LLVIP融合图像的平均质量评估指标

	EN	MI	SD	AG	VIF	SCD	MS_SSIM
DenseFuse	6.9740	13.9480	36.7708	2.8641	0.4617	1.3835	0.9224
RFN-Nest	7.0028	14.0055	37.8919	2.2445	0.4216	1.4154	0.8981
FusionGAN	6.4153	12.8306	26.1540	2.0102	0.2746	0.7521	0.7865
U2fusion	6.6531	13.3062	34.9659	3.3391	0.5283	1.2755	0.9098
CSF	6.8283	13.6566	35.3773	2.7960	0.4564	1.3621	0.9109
SEDR	6.8877	13.7795	36.5430	2.6319	0.4495	1.2546	0.8834
swinfusion	7.3844	14.7688	50.8104	**4.2064**	0.9087	1.5887	**0.9451**
seAFusion	7.4193	14.8386	50.4468	4.1898	0.9062	1.6259	0.9435
CDDFuse	7.3134	14.6267	48.3450	3.8052	0.8171	1.5889	0.9337
ours	**7.5398**	**15.0795**	**52.9415**	3.8464	**0.9889**	**1.7071**	0.9250

References

[1] Xu H, Tian X, *et al*. Image fusion meets deep learning: A survey and perspective[J]. *Information Fusion*, 2021, **76**: 323-336.

[2] Li J, Liu J, Zhou S, *et al*. Infrared and visible image fusion based on residual dense network and gradient loss[J]. *Infrared Physics & Technology*, 2023, **128**: 104486.

[3] Meher B, Agrawal S, Panda R, *et al*. A survey on region based image fusion methods[J]. *Information Fusion*, 2019, **48**: 119-132.

[4] Li S, Kang X, Fang L, *et al*. Pixel-level image fusion: A survey of the state of the art[J]. *information Fusion*, 2017, **33**: 100-112.

[5] Ma J, Chen C, Li C, *et al*. Infrared and visible image fusion via gradient transfer and total variation minimization[J]. *Information Fusion*, 2016, **31**: 100-109.

[6] Chen J, Li X, Luo L, *et al*. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition[J]. *Information Sciences*, 2020, **508**: 64-78.

[7] Li J, Liu J, Zhou S, *et al*. Learning a coordinated network for detail-refinement multiexposure image fusion[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **33**(2): 713-727.

[8] Li H, Wu X J, Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. *Information Fusion*, 2021, **73**: 72-86.

[9] Long Y, Jia H, Zhong Y, *et al*. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion[J]. *Information Fusion*, 2021, **69**: 128-141.

[10] Xu H, Liang P, Yu W, *et al*. Learning a Generative Model for Fusing Infrared and Visible Images via Conditional Generative Adversarial Network with Dual Discriminators[C]//IJCAI. 2019: 3954-3960.

[11] LIN Z P, LUO Y H, LI B Y, *et al*. Gradient-aware channel attention network for infrared small target image denoising before detection[J]. *Journal of Infrared and Millimeter Waves*(林再平,罗伊杭,李博扬,等.基于梯度可感知通道注意力模块的红外小目标检测前去噪网络[J].红外与毫米波学报), 2024, **43**(2): 254-260.

[12] Li B, Wang Y, Wang L, *et al*. Monte Carlo linear clustering with single-point supervision is enough for infrared small target detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 1009-1019.

[13] Li B, Wang L, Wang Y, *et al*. Mixed-Precision Network Quantization for Infrared Small Target Segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[14] Liu T, Yang J, Li B, *et al*. Infrared small target detection via nonconvex tensor tucker decomposition with factor prior[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[15] Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images[J]. *IEEE Transactions on Image Processing*, 2018, **28**(5): 2614-2623.

[16] Li Q, Han G, Liu P, *et al*. A multilevel hybrid transmission network for infrared and visible image fusion[J]. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: 1-14.

[17] Tang L, Xiang X, Zhang H, *et al*. DIVFusion: Darkness-free infrared and visible image fusion[J]. *Information Fusion*, 2023, **91**: 477-493.

[18] Zhao Z, Bai H, Zhang J, *et al*. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5906-5916.

[19] Tang L, Yuan J, Ma J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network[J]. *Information Fusion*, 2022, **82**: 28-42.

[20] Piella G. A general framework for multiresolution image fusion: from pixels to regions[J]. *Information fusion*, 2003, **4**(4): 259-280.

[21] Sadjadi F. Comparative image fusion analysais[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops. IEEE, 2005: 8-8.

[22] Young R K. Wavelet theory and its applications[M]. Springer Science & Business Media, 2012.

[23] Da Cunha A L, Zhou J, Do M N. The nonsubsampled contourlet transform: theory, design, and applications[J]. *IEEE transactions on image processing*, 2006, **15**(10): 3089-3101.

[24] LIN Z P, LI B Y, LI M ,*et al*. Light-weight infrared small target detection combining cross-scale feature fusion with bottleneck attention module [J]. *Journal of Infrared and Millimeter Wave*(林再平,李博扬,李淼,等.结合跨尺度特征融合与瓶颈注意力模块的轻量型红外小目标检测网络 [J].红外与毫米波学报), 2022, **41**(06): 1102-1112.

[25] Jian L, Yang X, Liu Z, *et al*. SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion [J]. *IEEE Transactions on Instrumentation and Measurement*, 2020, **70**: 1-15.

[26] Wang Z, Wu Y, Wang J, *et al*. Res2Fusion: Infrared and visible image fusion based on dense Res2net and double nonlocal attention

models［J］. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: 1-12.

［27］Zhao Z, Xu S, Zhang J, *et al*. Efficient and model-based infrared and visible image fusion via algorithm unrolling［J］. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **32**(3): 1186-1196.

［28］Woo S, Park J, Lee J Y, *et al*. Cbam: Convolutional block attention module［C］//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

［29］Zhao Y, Lv W, Xu S, *et al*. Detrs beat yolos on real-time object detection［C］//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16965-16974.

［30］Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection［J］. *arXiv preprint arXiv*:1911.09516, 2019.

［31］Xu H, Ma J, Le Z, *et al*. Fusiondn: A unified densely connected network for image fusion［C］//Proceedings of the AAAI conference on artificial intelligence. 2020, **34**(07): 12484-12491.

［32］Jia X, Zhu C, Li M, *et al*. LLVIP: A visible-infrared paired dataset for low-light vision［C］//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3496-3504.

［33］Ma J, Yu W, Liang P, *et al*. FusionGAN: A generative adversarial network for infrared and visible image fusion［J］. *Information fusion*, 2019, **48**: 11-26.

［34］Xu H, Ma J, Jiang J, *et al*. U2Fusion: A unified unsupervised image fusion network［J］. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **44**(1): 502-518.

［35］Xu H, Zhang H, Ma J. Classification saliency-based rule for visible and infrared image fusion［J］. *IEEE Transactions on Computational Imaging*, 2021, **7**: 824-836.

［36］Ma J, Tang L, Fan F, *et al*. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer［J］. *IEEE/CAA Journal of Automatica Sinica*, 2022, **9**(7): 1200-1217.

［37］Roberts J W, Van Aardt J A, Ahmed F B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification［J］. *Journal of Applied Remote Sensing*, 2008, **2**(1): 023522.

［38］Qu G, Zhang D, Yan P. Information measure for performance of image fusion［J］. *Electronics letters*, 2002, **38**(7): 1.

［39］Rao Y J. In-fibre Bragg grating sensors［J］. *Measurement science and technology*, 1997, **8**(4): 355.

［40］Cui G, Feng H, Xu Z, *et al*. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition［J］. *Optics Communications*, 2015, **341**: 199-209.

［41］Han Y, Cai Y, Cao Y, *et al*. A new image fusion performance metric based on visual information fidelity［J］. *Information fusion*, 2013, **14**(2): 127-135.

［42］Aslantas V, Bendes E. A new image quality metric for image fusion: The sum of the correlations of differences［J］. *Aeu-international Journal of electronics and communications*, 2015, **69**(12): 1890-1896.

［43］Ma K, Zeng K, Wang Z. Perceptual quality assessment for multi-exposure image fusion［J］. *IEEE Transactions on Image Processing*, 2015, **24**(11): 3345-3356.