

Infrared and visible image fusion based on transformer background modeling and CAM detail enhancement

Ji Xiao-Jian¹, Yu Chun-Yu¹, Chen Lu-Jie¹, Zhang Jun-Ju², Sun Bin³

- (1. College of Electronic and Optical Engineering & College of Flexible Electronics (Future Technology), Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. College of Electronic Engineering and Optoelectronic Technology, Nanjing University of Science and Technology, Nanjing 210094, China;
3. College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: To address the problems of insufficient background structure modeling and inadequate detail texture representation in infrared and visible image fusion, a dual-branch image fusion network based on Swin Transformer Background Modeling and Residual Channel Attention Detail Enhancement (ST-RCAFuse) was proposed. The specific design method was as follows. In the encoding stage, two branches, namely a background branch and a detail branch, were designed. In the background branch, the Swin Transformer was adopted as the core component. The Window-based Self-Attention (WSA) mechanism was used to achieve efficient modeling of both global and local background structures. Coordinate Attention (CoordAtt) was introduced to enhance the spatial directionality of features. In the detail branch, the Residual Channel Attention Block (RCAB) was employed to extract texture details and high-frequency information. In the decoding stage, the background and detail features were progressively fused and reconstructed to generate high-quality fused images. The FLIR dataset was selected for network training, and a multi-dimensional evaluation framework consisting of in-domain, conventional out-of-domain, and extreme scenarios was established. Specifically, the FLIR test set was used as the in-domain test group to evaluate the baseline performance under the same data distribution. The TNO and RoadScene datasets were adopted as conventional out-of-domain test groups to assess cross-scene generalization capability. In addition, two extreme scenario test sets, including a nighttime strong illumination interference dataset and an ultra-low-light field dataset, were constructed to comprehensively evaluate the robustness of the model under complex and adverse conditions. The experimental results demonstrate that the proposed ST-RCAFuse achieves superior visual quality on standard public datasets, with leading performance in key metrics such as entropy, spatial frequency, standard deviation, and average gradient. Furthermore, under extreme conditions such as nighttime strong illumination interference and ultra-low-light field environments, the method effectively suppresses interference, preserves fine details and enhances salient targets. The fusion performance is significantly better than that of existing comparison methods, fully validating its excellent generalization capability, robustness, and practical value across diverse scenarios and challenging conditions.

Key words: dual-branch, image fusion, two-stage training, coordinate attention, residual channel attention

Introduction

Infrared and visible image fusion (IVIF) is an important research topic in the field of image fusion. It aims to generate more informative images by integrating complementary information from infrared and visible images^[1]. Infrared images can effectively avoid the interference caused by illumination variations and artifacts on visual perception, but they usually suffer from low spatial resolution and weak texture details. Visible images, on the other hand, contain rich gradient information and

high spatial resolution, but they are easily affected by illumination conditions and occlusion. Therefore, the fused images can combine the advantages of both imaging modalities and have been widely applied in security monitoring^[2], image enhancement^[3], target tracking^[4], and object detection^[5]. With the rapid development of artificial intelligence, image fusion methods have gradually evolved from traditional methods to deep learning-based methods.

Traditional image fusion methods mainly include

Received date: 2026-03-24,

收稿日期: 2026-03-24, **录用日期:** 2026-04-24

Foundation items: Supported by the Key Laboratory Fund of Yunnan Province, High-Quality Generation Network from Visible Images to Infrared and Low-Light Images (2025-LLDIVN-GD-01-06).

multi-scale transform methods^[6], hybrid methods^[7], saliency-based methods^[8], subspace-based methods^[9], and sparse representation methods^[10]. Most of these methods rely on manually designed features and fusion rules, which have poor adaptability to complex scenes and are difficult to fully exploit the complementary information between infrared and visible images. Therefore, researchers have gradually introduced neural networks into IVIF tasks and proposed a series of fusion methods based on deep learning (DL). In 2019, Li et al. proposed the DenseFuse method^[11], which introduced dense connection structures into an encoder-decoder framework to achieve efficient feature fusion. In 2020, Li et al. proposed the NestFuse method^[12], which extracts multi-scale features through a nested encoder-decoder architecture and incorporates spatial and channel attention mechanisms, effectively enhancing the preservation of background details and salient target information. In 2021, Li et al. proposed RFN-Nest^[13], which further introduced a residual fusion network to achieve end-to-end fusion modeling. In 2022, Tang et al. proposed the semantic-aware fusion network SeAFusion^[14], which introduces semantic loss to constrain high-level feature feedback, highlighting infrared targets while preserving visible texture information. In 2023, Zhao et al. proposed a correlation-driven dual-branch feature decomposition fusion network, CDDFuse^[15], which employs a lite Transformer and an invertible neural network to model global low-frequency and local high-frequency features, respectively. The method achieves excellent performance in multimodal fusion tasks, including infrared - visible and medical image fusion, and significantly improves the accuracy of downstream tasks such as detection and segmentation. In 2024, Li et al. proposed CrossFuse^[16], which strengthens inter-modal information interaction through a cross-attention mechanism and reduces redundant information in fused results. In the same year, Tang et al. proposed ITFuse^[17], which constructs an interactive Transformer-based fusion framework to effectively capture the commonality and differences of multimodal features. In 2025, Cheng et al. proposed LEFuse^[18], which jointly models low-light enhancement and fusion tasks and achieves collaborative extraction of global and local information through a hybrid network architecture, significantly improving fusion image quality in nighttime scenes. Also in 2025, Song et al. proposed SFINet^[19], which combines image fusion with semantic segmentation tasks and realizes feature interaction through multi-scale feature extraction and attention mechanisms, balancing visual quality and semantic representation in all-day scenarios.

Although the above methods have continuously improved feature extraction and fusion strategies, there is still a lack of balance between global semantic modeling and local high-frequency detail preservation. To address this issue, this paper proposes a dual-branch collaborative image fusion network, namely Swin Transformer Background Modeling and Residual Channel Attention Detail Enhancement for Image Fusion (ST-RCAFuse),

which leverages the strengths of Convolutional Neural Networks (CNNs) in local detail extraction and Transformer^[20] in capturing global dependencies via self-attention mechanisms.

1 Working principle of ST-RCAFuse

1.1 Network structure of ST-RCAFuse

The ST-RCAFuse training framework adopts a two-stage training strategy and consists of three main components: a dual-branch encoder for feature extraction and decomposition, a decoder used to reconstruct the source images in Training stage I or generate fused images in Training stage II, and a background/detail fusion layer for fusing features with different frequency characteristics. The dual-branch encoder is composed of three parts: a shared encoder consisting of two shallow convolutional layers (Conv1 and Conv2), a Background Extraction Module (BEM) for background modeling, and a Detail Extraction Module (DEM) for detail modeling.

As illustrated in Fig. 1(a), Training stage I employs a self-supervised training strategy for feature decomposition, aiming to enable the encoder and decoder to learn stable decomposition of infrared or visible features and reconstruct the source images without distortion. Specifically, the source images are first fed into a shared feature encoder composed of shallow convolutional layers. The resulting shared feature $F_{shared} \in R^{C \times H \times W}$ is then passed through the Background Extraction Module (BEM) and the Detail Extraction Module (DEM) to generate the corresponding background and detail features, respectively. These two types of features are concatenated along the channel dimension and subsequently fed into a multi-level feature reconstruction decoder to produce the reconstructed image.

As illustrated in Fig. 1(b), Training stage II is performed based on the pre-trained encoder from Training stage I, with an additional fusion layer incorporated during training, so as to generate high-quality fused images with both structural consistency and fine detail clarity. The processing procedure is as follows:

First, the infrared I and visible images V are fed into the encoder $E(\cdot)$ to extract background features and detail features, respectively:

$$(\Phi_{IB}, \Phi_{ID}) = E(I), \quad (1)$$

$$(\Phi_{VB}, \Phi_{VD}) = E(V), \quad (2)$$

Where Φ_{IB} and Φ_{ID} represent the background and detail features decomposed from the infrared image I , respectively, and Φ_{VB} and Φ_{VD} represent those from the visible image V , respectively.

Subsequently, the corresponding features are fused via the background fusion layer $F_B(\cdot)$ and the detail fusion layer $F_D(\cdot)$:

$$\Phi_B = F_B(\Phi_{IB}, \Phi_{VB}), \quad (3)$$

$$\Phi_D = F_D(\Phi_{ID}, \Phi_{VD}), \quad (4)$$

Where Φ_B and Φ_D represent the fused background features and fused detail features, respectively.

Finally, the fused background features and detail

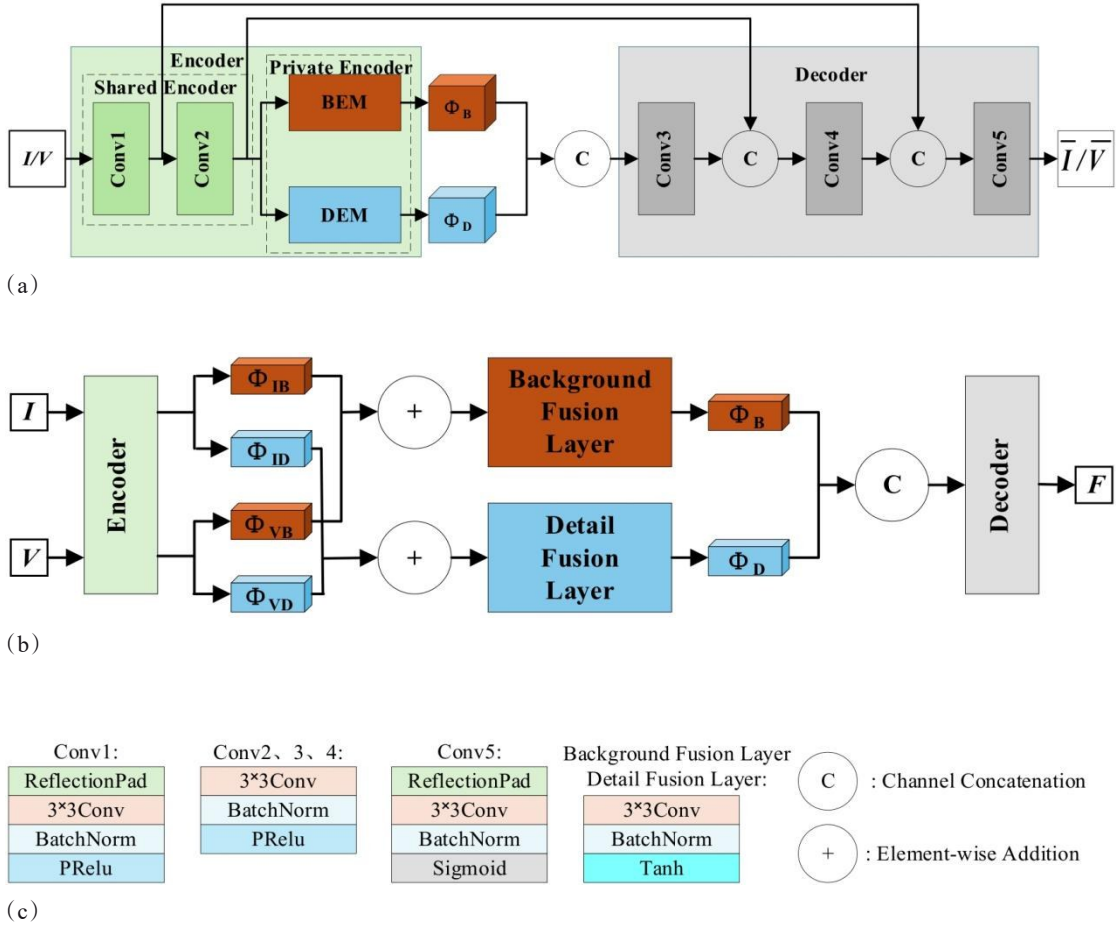


Fig. 1 Training network framework of ST-RCAFuse: (a)Training stage I; (b)Training stage II; (c)Network components and symbol description

图1 ST-RCAFuse 训练网络框架: (a)训练阶段 I; (b)训练阶段 II; (c)网络组成与符号说明

features are concatenated along the channel dimension via the operation $\text{Concat}(\cdot)$, and subsequently fed into the decoder $D(\cdot)$ to produce the final fused image F :

$$F = D(\text{Concat}(\Phi_B, \Phi_D)), \quad (5)$$

1.2 Background Extraction Module (BEM)

The Background Extraction Module (BEM) aims to extract low-frequency structural information from infrared and visible images and improve the consistency of cross-modal background representation. As illustrated in Fig. 2, the module first introduces the Coordinate Attention (CoordAtt) mechanism, which enhances the spatial perception capability of features in the horizontal and vertical directions via decomposed coordinate encoding, thereby emphasizing stable and continuous regional structural features. Subsequently, two layers of Swin Transformer^[21] are employed as the core component for background modeling. The local self-attention based on window partition is used to model local features, while the shifted window mechanism enables cross-window feature interaction, thereby progressively aggregating global contextual information. This allows the network to capture large-scale semantic structures, homogeneous back-

ground regions, and object contour information in the scene. Through the collaborative effect of these components, the module finally outputs background feature representations with high consistency, low noise, and stable semantic information.

1.2.1 Coordinate Attention (CoordAtt)

The core idea of CoordAtt is to decompose channel attention (CA) into horizontal (W) and vertical (H) directions while preserving spatial information. The specific implementation process is illustrated in Fig. 3.

For the input feature $F_{shared} \in R^{C \times H \times W}$ extracted by the shared encoder, horizontal pooling and vertical pooling are performed respectively, to obtain:

$$F_{shared}^h(c, i, 1) = \frac{1}{W} \sum_{j=1}^W F_{shared}(c, i, j), \quad (6)$$

$$F_{shared}^w(c, 1, j) = \frac{1}{H} \sum_{i=1}^H F_{shared}(c, i, j), \quad (7)$$

Where $F_{shared}^h \in R^{C \times H \times 1}$ and $F_{shared}^w \in R^{C \times 1 \times W}$. Then, the pooled features from the two directions are concatenated to obtain Y :

$$Y = \text{Concat}(F_{shared}^h, F_{shared}^w), \quad (8)$$

Where $Y \in R^{C \times (H+W) \times 1}$. Then, the concatenated feature is subjected to channel reduction and nonlinear transformation to obtain Z :

$$Z = \delta(\text{BN}(W_1 * Y)), \quad (9)$$

Where W_1 denotes the weight of the 1×1 convolution, which reduces the number of channels from C to C/R , and $\delta(\cdot)$ represents the h-swish activation function. Subsequently, Z is split into vertical and horizontal directional features, denoted as Z_h and Z_w :

$$Z_h = Z[:, :, 0:H, :], \quad Z_w = Z[:, :, H::, :], \quad (10)$$

Where $Z_h \in R^{C/R \times H \times 1}$ and $Z_w \in R^{C/R \times 1 \times W}$, then two attention weights are generated:

$$A_h = \sigma(W_h * Z_h), \quad A_w = \sigma(W_w * Z_w), \quad (11)$$

Where W_h and W_w denote the weights of the 1×1 convolutions, which finally output features with C channels, σ represents the Sigmoid activation function, $A_h \in R^{C \times H \times 1}$ is the horizontal attention weight, and $A_w \in R^{C \times 1 \times W}$ is the vertical attention weight. The final attention is fused to obtain the output:

$$F_{ca} = F_{shared} \times A_h \times A_w, \quad (12)$$

Where $F_{ca} \in R^{C \times H \times W}$.

1.2.2 Architecture of Swin Transformer

The architecture of the Swin Transformer is illustrated

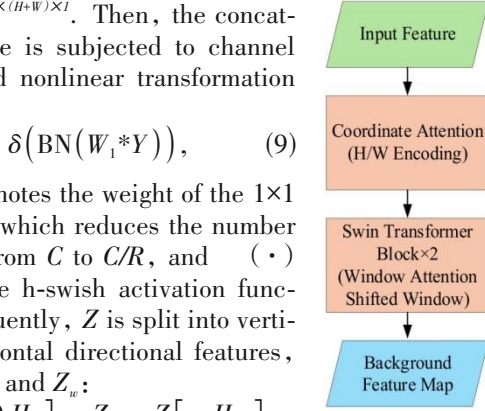


Fig. 2 Architecture of the Background Extraction Module
图2 背景提取模块结构图

in Fig. 4, which takes the output F_{ca} of the CoordAtt module described in Section 1.2.1 as its input. The processing procedure is as follows.

The input F_{ca} first undergoes channel normalization via *LayerNorm*:

$$F_{norm} = \text{LayerNorm}(F_{ca}), \quad (13)$$

Where *LayerNorm*(\cdot) denotes the layer normalization operator, which operates along the channel dimension of the features to stabilize the training and accelerate model convergence.

The normalized features are partitioned into non-overlapping windows:

$$F_{windows} = \text{WindowPartition}(F_{norm}), \quad (14)$$

Where *WindowPartition*(\cdot) denotes the window partitioning operator, which splits the input feature map into non-overlapping local windows to prepare for the subsequent local self-attention computation.

Multi-head Self-Attention (MSA) is performed on each window:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B_{rel}\right)V, \quad (15)$$

Where Q , K , and V are the query, key, and value matrices obtained via linear transformation, respectively, d denotes the dimension of each attention head, B_{rel} is the relative position bias used to encode the positional information within the window, and *Softmax*(\cdot) is the activation function applied to normalize the attention weights.

The attention output is restored to the original feature map size via the inverse window partitioning operation, and then added to the input through a residual con-

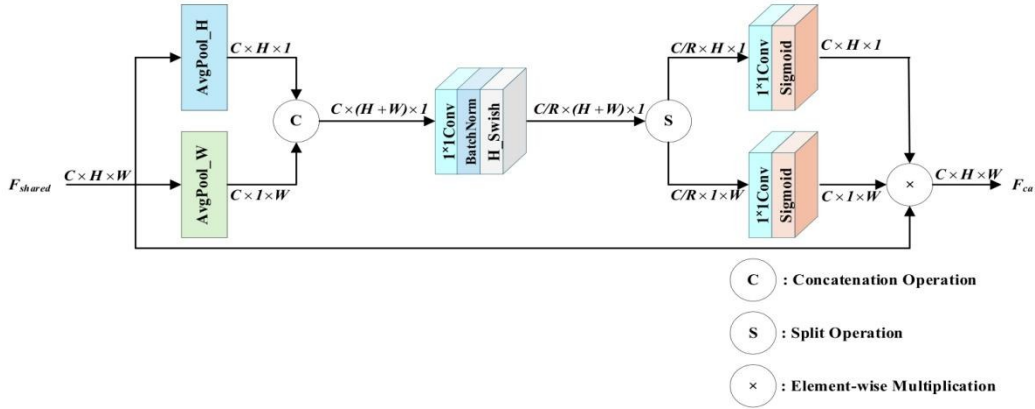


Fig. 3 Network architecture of CoordAtt
图3 CoordAtt网络结构

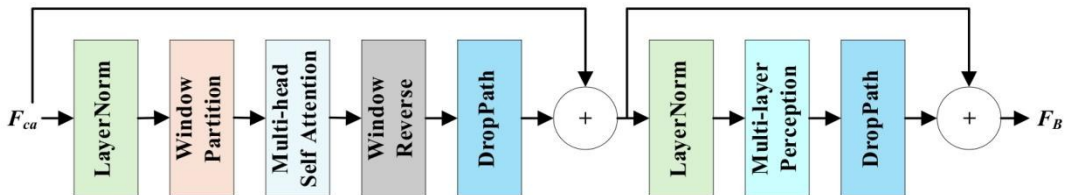


Fig. 4 Network architecture of the Swin Transformer
图4 Swin Transformer网络结构

nection to obtain F_{attn} :

$$F_{attn} = F_{ca} +$$

$$DropPath\left(WindowReverse\left(Attention\left(F_{windows}\right)\right)\right), (16)$$

Where $WindowReverse(\cdot)$ is the inverse window partitioning operator, which acts oppositely to $WindowPartition(\cdot)$ and is used to restore the windowed features to the original feature map size; $DropPath(\cdot)$ is the stochastic depth regularization operator, which randomly drops the residual paths with a certain probability to prevent model overfitting; and $Attention(\cdot)$ denotes the multi-head self-attention computation defined in Equation (15).

The attention output is further processed by $LayerNorm(\cdot)$, an $MLP(\cdot)$ mapping, and a residual connection to obtain the final output F_B :

$$F_B = F_{attn} + DropPath\left(MLP\left(LayerNorm\left(F_{attn}\right)\right)\right), (17)$$

Where $MLP(\cdot)$ denotes the multi-layer perceptron operator, which is used to perform non-linear feature transformation.

1.3 Detail Extraction Module (DEM)

The Detail Extraction Module (DEM) employs a Residual Channel Attention Block (RCAB). The RCAB is composed of two 3×3 convolution layers, a Channel Attention (CA) mechanism, and a residual connection. This structure enables the extraction of local texture information while emphasizing salient detail regions through channel reweighting, thereby enhancing feature representation capability.

As shown in Fig. 5, the input features $F_{shared} \in R^{C \times H \times W}$ extracted by the shared encoder are first passed through a 3×3 convolution layer followed by a ReLU activation function to extract basic texture features. Then, a second 3×3 convolution layer is applied to further refine local structural information. On this basis, a channel attention mechanism is introduced to adaptively reweight the feature responses. The channel attention module obtains global contextual information from convolutional features through global average pooling. Then, two channel-wise 1×1 convolution layers are used to model nonlinear channel dependencies and feature interactions, followed by a Sigmoid function to generate channel attention weights. These weights are used to modulate the features obtained from the second convolution layer in a channel-wise manner, thereby enhancing channels containing important texture information while suppressing redundant or noisy features. Finally, a residual connection is employed to add the modulated features to the input features to pro-

duce the final output feature F_D , allowing the network to emphasize detailed information while ensuring stable convergence.

1.4 Loss Function

1.4.1 Loss function for Training Stage I

In Training Stage I, the model employs a self-supervised approach to perform feature decomposition on infrared or visible images. The goal is to enable the encoder to stably extract and distinguish between background features and detail features, while allowing the decoder to reliably reconstruct the input images. To this end, a composite loss function consisting of reconstruction loss, feature decomposition loss, and gradient consistency loss is adopted in this stage. The overall loss can be expressed as:

$$L^1 = L_{rec}^{vis} + L_{rec}^{ir} + \alpha_1 L_{decomp}^1 + \beta_1 L_{grad}^1, (18)$$

Where L_{rec}^{vis} denotes the visible image reconstruction loss, L_{rec}^{ir} represents the infrared image reconstruction loss, L_{decomp}^1 is the feature decomposition loss, and L_{grad}^1 stands for the gradient consistency loss.

The reconstruction loss L_{rec} is used to constrain the decoder output to be consistent with the original input, enabling the network to acquire reliable basic reconstruction capability. It combines the Structural Similarity Index Measure (SSIM) and the Mean Squared Error (MSE) to balance structural fidelity and pixel-level accuracy:

$$L_{rec}^{vis} = 5 \cdot SSIM\left(I_{vis}, \hat{I}_{vis}\right) + \left\|I_{vis} - \hat{I}_{vis}\right\|_2^2, (19)$$

$$L_{rec}^{ir} = 5 \cdot SSIM\left(I_{ir}, \hat{I}_{ir}\right) + \left\|I_{ir} - \hat{I}_{ir}\right\|_2^2, (20)$$

Where I_{ir} and I_{vis} denote the input infrared and visible images, respectively, and \hat{I}_{ir} and \hat{I}_{vis} represent the corresponding reconstruction results of infrared and visible images.

The feature decomposition loss L_{decomp}^1 is based on the Correlation Constraint (CC). This loss enhances the consistency between cross-modal background features while reducing the correlation between detail features, thereby guiding the network to achieve effective decoupling of background and detail information.

$$L_{decomp}^1 = \frac{\left(CC\left(\Phi_{ir}^D, \Phi_{vis}^D\right)\right)^2}{1.01 + CC\left(\Phi_{ir}^B, \Phi_{vis}^B\right)}, (21)$$

Where Φ^D and Φ^B represent the detail features and background features of the two-modal inputs, respectively.

The gradient consistency loss L_{grad}^1 is the L1 norm of

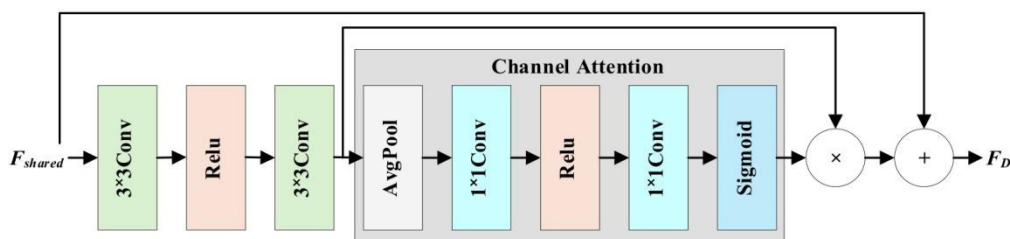


Fig. 5 Residual Channel Attention Block (RCAB)
图5 残差通道注意力模块(RCAB)

the Sobel gradient difference between the input image and the reconstructed image, which is used to improve the edge and texture fidelity of the reconstruction results.

$$L_{\text{grad}}^I = \left\| \nabla I - \nabla \hat{I} \right\|_1, \quad (22)$$

1.4.2 Loss function for Training Stage II

In Training Stage II, the task shifts from self-supervised reconstruction to infrared-visible feature fusion, which integrates the thermal radiation information from infrared images and the texture details from visible images to generate a fused image with richer information content. Therefore, a composite loss function consisting of fusion loss and feature decomposition loss is adopted in this stage. The loss function is defined as follows:

$$L^{\text{II}} = L_{\text{fusion}}^{\text{II}} + \alpha_2 L_{\text{decomp}}^{\text{II}}, \quad (23)$$

Where $L_{\text{fusion}}^{\text{II}}$ denotes the fusion loss, and $L_{\text{decomp}}^{\text{II}}$ represents the feature decomposition loss.

The fusion loss $L_{\text{fusion}}^{\text{II}}$ is used to constrain the generated image to contain the key information of both visible and infrared images, including intensity information at the pixel level and structural details at the gradient level.

The intensity fidelity loss $L_{\text{intensity}}^{\text{II}}$ is designed to ensure that the generated image retains the salient regions of both modalities at the pixel level, adopting a maximum response strategy:

$$L_{\text{intensity}}^{\text{II}} = \left\| \max(I_{\text{vis}}^Y, I_{\text{ir}}) - I_f \right\|_1, \quad (24)$$

Its physical meaning is that the fused image I_f generated by the network should approach the regions with higher brightness in the two modalities, so as to preserve the main information of infrared targets and visible light high-brightness regions.

The gradient fidelity loss is designed to further preserve the texture details in visible images and the contour structures in infrared images:

$$I_{\text{grad}}^{\text{vis}} = \left| \nabla_x I_{\text{vis}}^Y \right| + \left| \nabla_y I_{\text{vis}}^Y \right|, \quad I_{\text{grad}}^{\text{ir}} = \left| \nabla_x I_{\text{ir}} \right| + \left| \nabla_y I_{\text{ir}} \right|, \quad (25)$$

$$I_{\text{grad}}^{\text{joint}} = \max(I_{\text{grad}}^{\text{vis}}, I_{\text{grad}}^{\text{ir}}), \quad (26)$$

$$L_{\text{grad}}^{\text{II}} = \left\| I_{\text{grad}}^{\text{joint}} - I_{\text{grad}}^f \right\|_1, \quad (27)$$

This term forces the fused image to cover the salient gradient features of both modalities in terms of edges, textures, and structural information simultaneously. The final fusion loss $L_{\text{fusion}}^{\text{II}}$ is expressed as:

$$L_{\text{fusion}}^{\text{II}} = L_{\text{intensity}}^{\text{II}} + 10L_{\text{grad}}^{\text{II}}, \quad (28)$$

The feature decomposition loss $L_{\text{decomp}}^{\text{II}}$ is identical to L_{decomp}^I in Stage I.

2 Experimental results and analysis

2.1 Experimental setting

All experiments are conducted on a Windows 11 operating system with an NVIDIA GeForce RTX 4060 GPU (8 GB memory). The implementation is based on Python 3.8 and the PyTorch deep learning framework. The two-stage network is trained for a total of 60 epochs, where Training Stage I and Stage II are set to 20 epochs and 40 epochs, respectively. The batch size is set to 24. The Adam optimizer is adopted with an initial learning rate of

10^{-4} , which is decayed to 0.1 times the original value every 20 epochs.

The FLIR ADAS dataset is originally designed for object detection tasks, containing tens of thousands of images with rich annotations. However, it includes a large number of highly redundant consecutive video frames and is not specifically tailored for image fusion tasks. To construct a high-quality training dataset suitable for fusion, 180 pairs of well-registered infrared-visible images were carefully selected from the original dataset. These image pairs are characterized by high visual quality, non-repetitive scenes, and cover a variety of typical scenarios, including urban roads, pedestrians, vehicles, and outdoor environments, ensuring the presence of salient thermal targets and clear visible texture details. Data augmentation techniques, including flipping, rotation, and random cropping, were applied to expand the training set to 720 image pairs. This dataset scale is sufficient to support effective training of the proposed dual-branch network, primarily due to the adoption of a two-stage training strategy, well-designed loss function constraints, and strong training stability. Before training, all images are converted to grayscale and center-cropped to a resolution of 128×128.

For the testing phase, a multi-dimensional evaluation framework is constructed, consisting of five test groups. First, 40 pairs of images from the FLIR test set are selected as the in-domain test group to evaluate the baseline performance under the same data distribution. Second, 40 image pairs from each of the TNO^[22] and RoadScene^[23] datasets are used as conventional out-of-domain test groups to assess the cross-scene generalization capability of the model. All three public datasets adopt widely used pre-registered versions in the literature. Furthermore, to comprehensively evaluate the robustness of the model under extreme conditions, two additional extreme scenario test sets are constructed. The first is a nighttime strong illumination interference dataset, which consists of 30 image pairs selected from the FLIR and RoadScene datasets, featuring typical challenging conditions such as intense road lighting, vehicle headlight glare, and building illumination. This dataset effectively evaluates the ability of fusion methods to handle highlight interference, overexposure, and texture suppression. The second is a self-built ultra-low-light field dataset introduced in this study. This dataset is collected in extremely low-illumination outdoor environments using a synchronized infrared thermal camera and a low-light camera, presenting dual challenges of extremely low illumination and weak target visibility. Eight representative image pairs are selected from video sequences as the test group. For this dataset, the SuperFusion^[24] registration network is employed to perform cross-modality alignment between infrared and visible images. This method effectively estimates spatial transformations between modalities, enabling high-precision alignment and providing high-quality registered images for subsequent fusion tasks. Overall, this “in-domain + conventional out-of-domain + extreme scenario” evaluation framework is de-

signed to comprehensively assess the fusion methods from multiple perspectives, including stability, generalization, and robustness.

2.2 Evaluation metrics for image fusion

In this paper, eight objective evaluation metrics are employed to quantitatively assess the fusion performance, including Entropy (EN)^[25], Spatial Frequency (SF)^[26], Standard Deviation (SD)^[27], Peak Signal-to-Noise Ratio (PSNR)^[28], Average Gradient (AG)^[29], Correlation Coefficient (CC)^[30], Sum of the Correlations of Differences (SCD)^[31], and Quality Assessment Based on Edge Information Preservation (Qabf)^[32]. Higher values of these metrics generally indicate better fusion image quality.

2.3 Comparative experiments with mainstream fusion methods

To verify the effectiveness of the proposed method, nine representative infrared and visible image fusion methods are selected for comparison, including DenseFuse^[11], NestFuse^[12], RFN-Nest^[13], SeAFusion^[14], CDDFuse^[15], CrossFuse^[16], ITFuse^[17], LEFuse^[18] and DSFusion^[33]. Comparative experiments are conducted on the aforementioned test datasets. Meanwhile, to further clarify the core differences between the proposed method and existing approaches, this section establishes a method comparison table (as shown in Table 1) from four key dimensions: network architecture, feature extraction, fusion strategy, and loss function design.

The performance of all fusion methods is comprehensively evaluated from both subjective visual quality and objective quantitative metrics. For subjective evaluation, infrared salient regions and visible detail regions in

each fused image are marked with red and green boxes, respectively. The corresponding Regions of Interest (ROI) are enlarged and placed below the original images to intuitively compare the performance of different methods in target saliency and detail preservation. For objective evaluation, statistical analysis is performed on the average results of the ten fusion methods across eight evaluation metrics. The best performance in each metric is highlighted in bold font, the second best is marked in red, and the third best is marked in green.

2.3.1 Analysis of fusion results on the FLIR test set

A representative pair of source images is selected from the FLIR test set and fused using different fusion methods for comparison, as shown in Fig. 6. The scene depicts a suburban residential street, containing large-scale objects such as vehicles, trees, and buildings, as well as fine details including manhole cover patterns and leaves. This scenario provides an effective evaluation of the fusion performance of different methods in real-world outdoor environments.

From the enlarged ROI regions in Fig. 6, it can be observed that the fusion images generated by DenseFuse and ITFuse exhibit relatively low overall contrast, and both the manhole cover patterns and leaf textures appear noticeably blurred. NestFuse and RFN-Nest fail to effectively balance infrared and visible information, resulting in fusion outputs that are overly biased toward the infrared modality, leading to abnormal darkening in the sky regions. SeAFusion and CDDFuse suffer from severe loss of fine details, particularly in the manhole cover patterns. CrossFuse preserves clear leaf textures but is un-

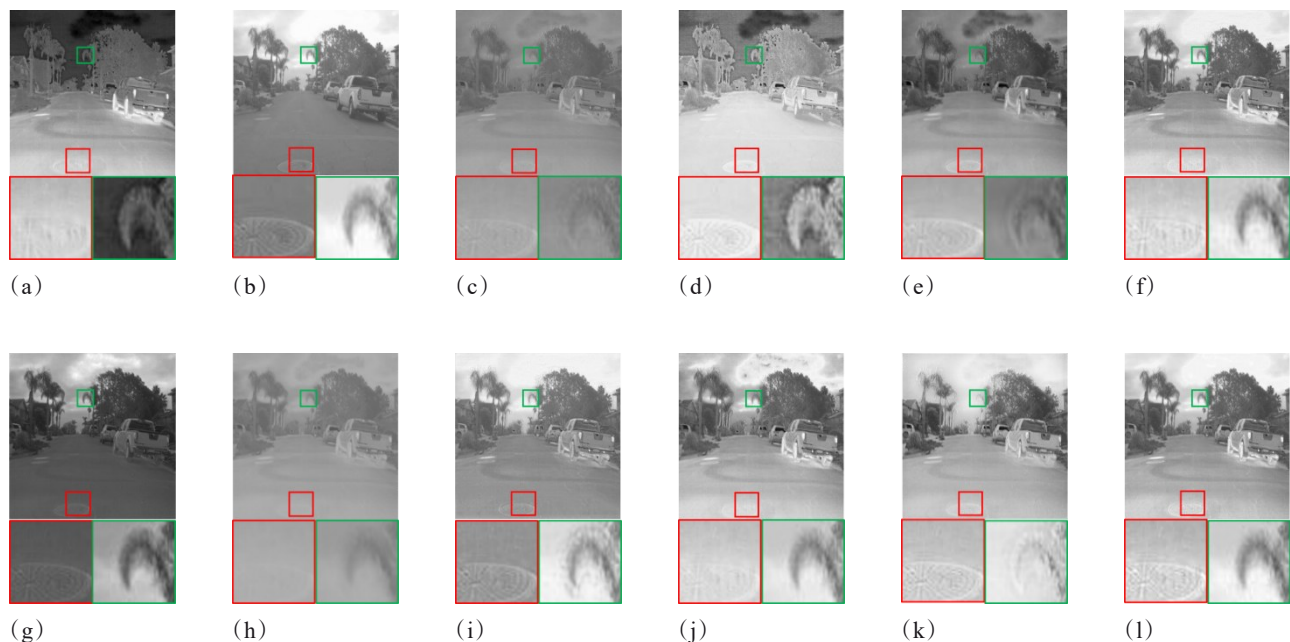


Fig. 6 Fusion results on the FLIR test set: (a) Infrared image; (b) Visible image; (c) DenseFuse; (d) NestFuse; (e) RFN-Nest; (f) SeAFusion; (g) CrossFuse; (h) ITFuse; (i) DSFusion; (j) CDDFuse; (k) LEFuse; (l) Proposed ST-RCAFuse

图6 FLIR测试集上的融合结果: (a)红外图像; (b)可见光图像; (c) DenseFuse; (d) NestFuse; (e) RFN-Nest; (f) SeAFusion; (g) CrossFuse; (h) ITFuse; (i) DSFusion; (j) CDDFuse; (k) LEFuse; (l) 本文 ST-RCAFuse

Table 1 Comparison of core designs for 10 image fusion methods**表1 10种图像融合方法的核心设计对比**

	Network Architecture	Feature Extraction	Fusion Strategy	Loss Function Design
DenseFuse	CNN-based encoder - decoder with dense blocks	Multi-level features extracted via convolution and dense connections	Direct addition or adaptive weighting based on the L_1 -norm	L_2 pixel loss and SSIM loss
NestFuse	Nested encoder - decoder with multi-scale structure	Multi-level deep features via multi-scale downsampling	Spatial and channel attention-based weighted fusion	Frobenius norm loss and SSIM loss
RFN-Nest	Nested architecture with residual fusion network	Multi-level features via multi-scale downsampling	Adaptive fusion via a learnable residual network	Stage I: pixel loss and SSIM loss; Stage II: detail preservation loss and feature enhancement loss
SeAFusion	Lightweight CNN with gradient residual dense blocks	Feature extraction via convolution and gradient residual dense blocks	Channel concatenation for efficient fusion	Content loss and semantic loss
CDDFuse	Dual-branch Transformer - CNN architecture	Global features via a lightweight Transformer; detail features via an invertible neural network	Separate fusion followed by channel concatenation and reconstruction	Reconstruction loss, correlation-driven decoupling loss, intensity loss, and gradient loss
CrossFuse	Dual-encoder with cross-modal attention Transformer	Modality features extracted via dense blocks and self-attention	Reverse softmax-based cross-attention to enhance complementarity	Stage I: pixel loss and SSIM loss; Stage II: intensity loss and gradient loss
ITFuse	Interactive Transformer architecture	Feature extraction via residual attention and interactive attention	Cross-modal attention with long-range Transformer fusion	Pixel loss and SSIM loss
LEFuse	U-Net with parallel Transformer - CNN structure	Global features via Restormer; local features via StarBlock	Multi-scale skip connections with channel-weighted fusion	Maximum entropy loss, perceptual loss, and texture loss
DSFusion	Encoder - decoder with local attention modules	Multi-branch local attention enhances detail extraction	Dual-branch feature concatenation with adaptive weighting	Target-aware pixel loss and contrastive texture loss
ST-RCA-Fuse	Dual-branch encoder - decoder with Swin Transformer and RCAB	Background branch models global structure; detail branch extracts texture features	Separate fusion of background and detail features followed by channel concatenation and progressive reconstruction	Stage I: reconstruction loss, feature decoupling loss, and gradient consistency loss; Stage II: fusion loss and feature decoupling loss

able to effectively highlight infrared salient targets such as vehicles. DSFusion and LEFuse exhibit noticeable noise in the leaf texture regions, compromising detail integrity. In contrast, the proposed ST-RCAFuse method not only preserves clear and complete leaf textures in the ROI regions but also effectively enhances salient thermal targets such as manhole patterns and vehicles. Meanwhile, the sky background maintains natural brightness, achieving a well-balanced and superior fusion of complementary information from both modalities.

Table 2 presents the objective evaluation results of

ten fusion methods on the FLIR test set. It can be observed that the proposed method achieves the best performance in five metrics, including EN, SF, SD, AG, and SCD, while ranking second in PSNR, CC, and Qabf, demonstrating stable and competitive overall performance. By jointly considering both subjective visual quality and objective evaluation results, the proposed ST-RCAFuse achieves an optimal balance between infrared thermal target enhancement and visible texture preservation on the in-domain FLIR test set, exhibiting stable performance and clear overall advantages.

Table 2 Quantitative evaluation of ten fusion methods on the FLIR test set**表2 FLIR测试集上10种融合方法的定量评价**

	EN	SF	SD	PSNR	AG	CC	SCD	Qabf
DenseFuse	6.6380	7.5114	28.5168	64.5287	2.8824	0.6325	1.4161	0.3833
NestFuse	7.3128	13.2497	53.0027	60.9171	4.8239	0.5118	1.4698	0.3900
RFN-Nest	7.1690	7.2587	40.8548	61.0896	2.9728	0.6221	1.7694	0.2984
SeAFusion	7.3552	15.3897	51.3183	61.6801	5.7844	0.5742	1.7112	0.5211
CrossFuse	7.1223	10.7287	45.3615	60.0340	3.8935	0.5263	1.3059	0.3687
ITFuse	6.3172	4.7570	24.7043	63.9625	1.9538	0.6222	1.2050	0.2256
DSFusion	6.9933	13.6714	39.8952	63.0621	5.1191	0.5421	1.3085	0.5620
CDDFuse	7.4153	14.0677	51.7415	62.2656	5.7031	0.6111	1.8693	0.5111
LEFuse	7.4292	15.1357	52.4222	61.2840	5.6661	0.5830	1.7022	0.3986
ST-RCAFuse	7.4301	16.3580	54.6607	64.1431	5.8363	0.6256	1.8871	0.5239

2.3.2 Analysis of fusion results on the TNO Dataset

Two pairs of source images are selected from the TNO dataset for fusion comparison using different methods, as shown in Figure 7. Scene 1 depicts a complex field environment, where the background contains texture interference such as trees and roads, and the target (pedestrian) exhibits low contrast against the background. This scenario can verify the capability of different methods to preserve targets and fuse details under low-contrast and complex texture background conditions. Scene 2 represents a typical urban environment containing buildings, vehicles, clouds, and pedestrians, which is used to assess the performance of different methods in terms of information complementarity and visual fidelity in multi-object scenes.

From the enlarged ROIs in Fig. 7, it can be observed that the fusion images generated by DenseFuse, ITFuse, and RFN-Nest exhibit relatively low overall contrast, and the infrared salient targets are not effectively highlighted; NestFuse introduces more noise during the fusion process, resulting in poor image clarity. As shown in Scene 1, some detail textures such as tree branches are obviously missing, and the infrared targets such as pedestrians are not sufficiently represented; Although CrossFuse preserves visible texture details to a certain extent, its ability to highlight infrared salient regions is lim-

ited; LEFuse suffers from severe detail texture loss, for example, in Scene 1, the texture of tree branches is significantly missing; In comparison, SeAFusion and DSFusion can enhance infrared salient targets and preserve part of the detailed information. However, they still show limitations in balancing visible and infrared information fusion. As shown in Scene 2, the structural information of clouds in the sky region is missing. Overall, the proposed ST-RCAFuse achieves better visual performance. It can effectively highlight infrared salient targets while preserving visible texture details more completely, achieving a better balance between the two types of information. For example, in Scene 1, the pedestrian target is clearly highlighted and the tree branch textures are well preserved. In Scene 2, the cloud structures in the sky region are better maintained.

Table 3 lists the objective evaluation results of ten fusion methods on the TNO dataset. From the comparison, it can be observed that the proposed ST-RCAFuse achieves the best values in five metrics, including EN, SF, SD, AG, and SCD. It ranks second in the Qabf metric and third in the PSNR and CC metrics. By comprehensively considering both subjective visual quality and objective evaluation results, the proposed ST-RCAFuse demonstrates outstanding performance on the TNO dataset in terms of information content, detail preservation, and gradient correlation.

Table 3 Quantitative evaluation of ten fusion methods on the TNO dataset**表3 TNO数据集上10种融合方法的定量评价**

	EN	SF	SD	PSNR	AG	CC	SCD	Qabf
DenseFuse	6.3518	6.3793	24.7829	64.1708	2.5148	0.5242	1.6056	0.3506
NestFuse	6.9781	13.8660	44.4994	61.2547	5.0132	0.4577	1.6000	0.3306
RFN-Nest	6.9908	5.8795	37.2490	62.1540	2.6822	0.5163	1.7979	0.3380
SeAFusion	7.1361	12.2526	44.3379	61.4195	5.0109	0.4763	1.7345	0.4862
CrossFuse	6.9397	9.9824	40.4506	61.4667	3.7912	0.3891	1.3427	0.4242
ITFuse	6.0530	3.9160	21.8012	63.5038	1.6871	0.5115	1.4713	0.2078
DSFusion	6.9459	11.1598	40.3329	62.3731	4.3880	0.4178	1.5645	0.5439
CDDFuse	7.0555	13.1046	44.5993	61.4474	5.1779	0.4921	1.8043	0.5157
LEFuse	7.1290	12.8621	44.3781	59.4632	4.8094	0.4365	1.5453	0.3265
ST-RCAFuse	7.1565	14.1998	45.4504	62.4230	5.3518	0.5145	1.8109	0.5352

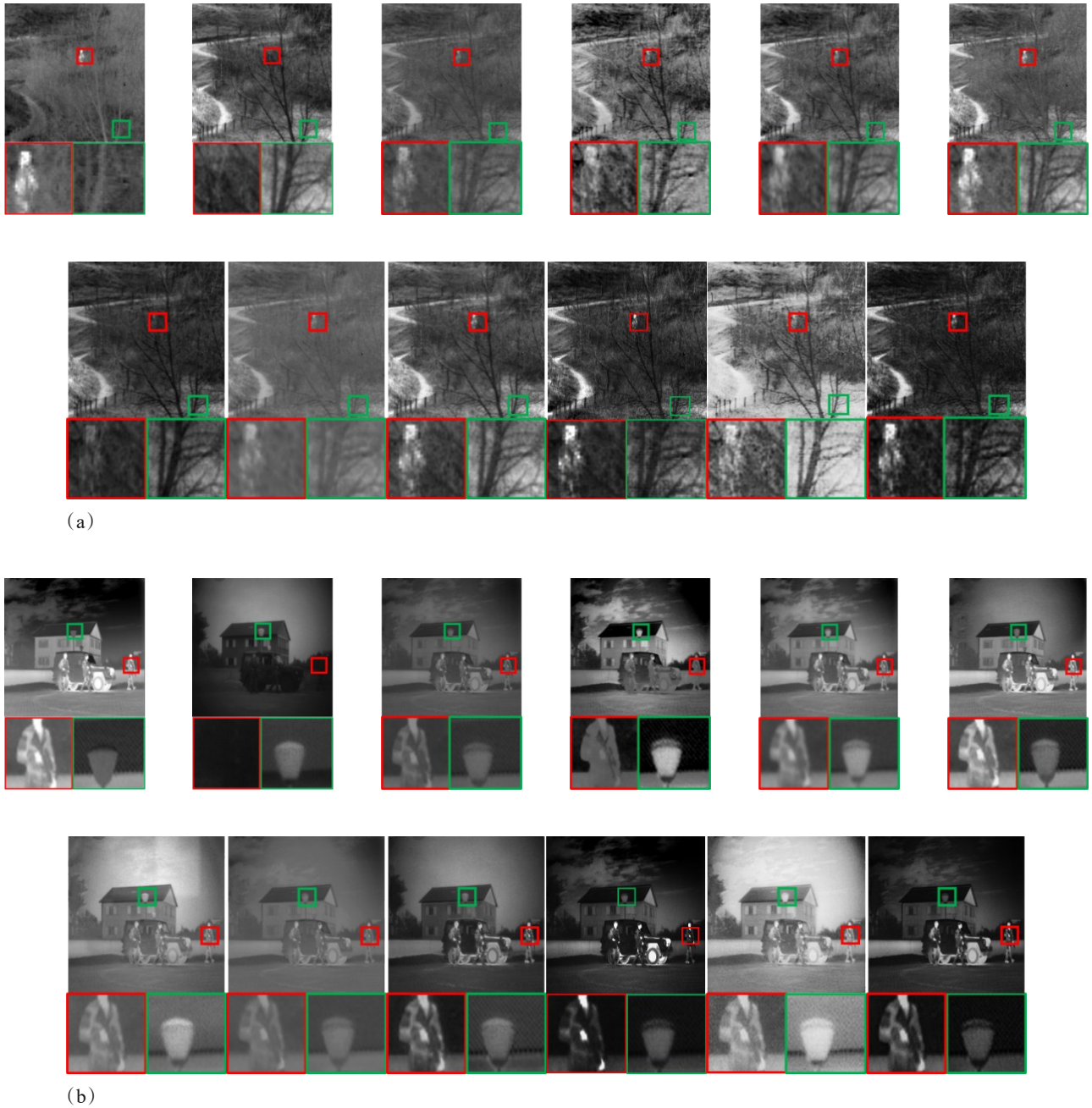


Fig. 7 Fusion results on the TNO dataset: (a) Scene 1; (b) Scene 2. In each scene, the sub-images from left to right and top to bottom are: Infrared image, Visible image, fusion results of DenseFuse, NestFuse, RFN-Nest, SeAFusion, CrossFuse, ITFuse, DSFusion, CDDFuse, LEFuse, and the proposed ST-RCAFuse.

图7 TNO数据集上的融合结果:(a)场景1;(b)场景2。每组场景内,子图从左到右、从上到下依次为:红外图像、可见光图像、DenseFuse、NestFuse、RFN-Nest、SeAFusion、CrossFuse、ITFuse、DSFusion、CDDFuse、LEFuse以及本文所提ST-RCAFuse的融合结果。

2.3.3 Analysis of fusion results on the RoadScene dataset

A representative pair of source images from a daytime urban road scene in the RoadScene dataset is selected for fusion comparison using different methods, as shown in Fig. 8. The scene includes pedestrians, a STOP traffic sign, buildings, and trees. The infrared image effectively highlights thermal targets such as pedestri-

ans and traffic signs, while the visible image retains rich environmental texture details, showing strong complementary characteristics between the two modalities.

From the enlarged ROI regions in Fig. 8, it can be observed that the fusion images generated by DenseFuse, ITFuse, and RFN-Nest exhibit relatively low overall contrast, and the infrared salient regions are not sufficiently highlighted. Meanwhile, detailed textures such as tree

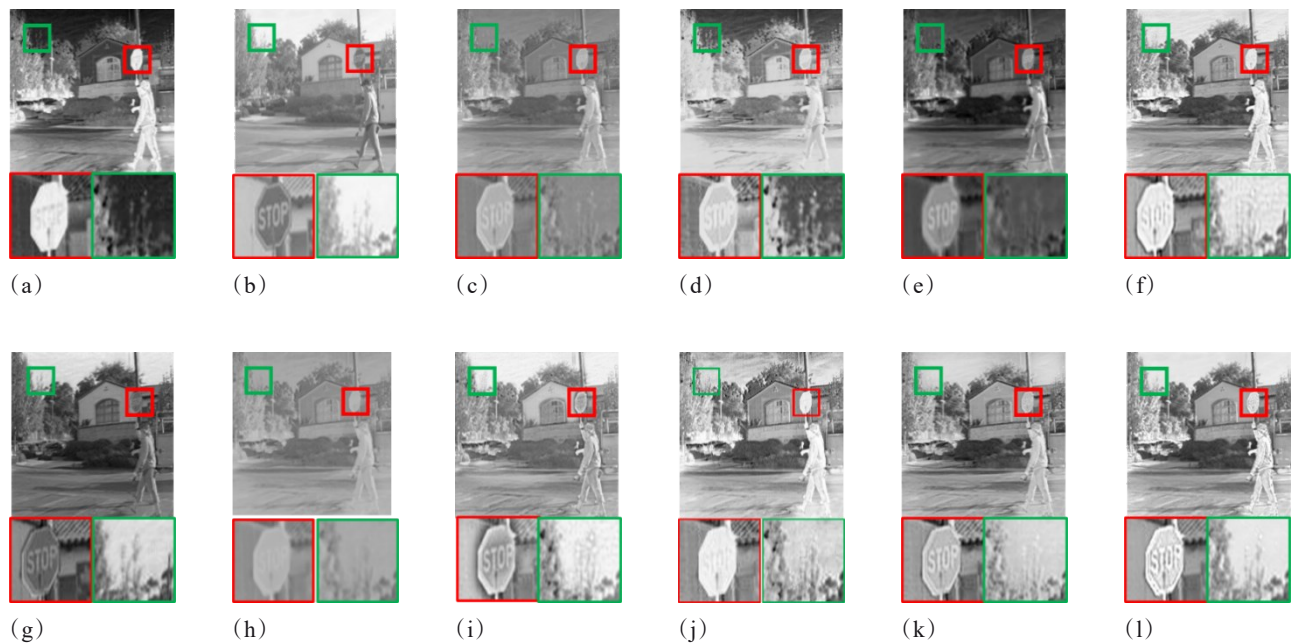


Fig. 8 Fusion results on the RoadScene dataset: (a) Infrared image; (b) Visible image; (c) DenseFuse; (d) NestFuse; (e) RFN-Nest; (f) SeAFusion; (g) CrossFuse; (h) ITFuse; (i) DSFusion; (j) CDDFuse; (k) LEFuse; (l) Proposed ST-RCAFuse

图8 RoadScene 数据集上的融合结果: (a)红外图像; (b)可见光图像; (c) DenseFuse; (d) NestFuse; (e) RFN-Nest; (f) SeAFusion; (g) CrossFuse; (h) ITFuse; (i) DSFusion; (j) CDDFuse; (k) LEFuse; (l)本文 ST-RCAFuse

branches appear blurred; NestFuse, SeAFusion, DSFusion, and LEFuse can highlight infrared salient targets to some extent, however, they are insufficient in preserving texture details and fail to clearly present fine structures such as tree branches; CrossFuse performs well in preserving texture details, but its ability to highlight infrared targets such as the traffic sign is limited; CDDFuse fails to clearly preserve the textual information on the traffic sign. In contrast, the proposed ST-RCAFuse can simultaneously enhance infrared salient information and visible texture details. In the fusion results, infrared targets are clearly highlighted, and texture structures such as tree branches are clearly distinguishable, resulting in a more balanced overall visual effect.

Table 4 presents the objective evaluation results of ten fusion methods on the RoadScene dataset. It can be

observed that the proposed method achieves the best performance in five metrics, namely EN, SF, SD, AG, and SCD, and ranks third in Qabf, PSNR, and CC, indicating stable and competitive overall performance. By jointly considering the subjective visual results and objective evaluation metrics, the proposed ST-RCAFuse demonstrates strong robustness on the RoadScene dataset.

2.3.4 Analysis of fusion results on the nighttime strong illumination interference dataset

A representative pair of source images is selected from the nighttime strong illumination interference dataset for fusion comparison, as shown in Fig. 9. The scene corresponds to a typical nighttime urban road environment. The visible image suffers from severe illumination interference, including direct vehicle headlights, strong traffic lights, and building illumination, resulting in

Table 4 Quantitative evaluation of ten fusion methods on the RoadScene dataset

表4 RoadScene数据集上10种融合方法的定量评价

	EN	SF	SD	PSNR	AG	CC	SCD	Qabf
DenseFuse	6.7866	8.4774	31.0295	64.5195	3.3601	0.6467	1.3938	0.3802
NestFuse	7.2400	15.1061	53.6296	60.8258	5.2894	0.4969	1.2436	0.3739
RFN-Nest	7.2992	7.8336	44.1396	61.2773	3.3894	0.6330	1.7273	0.2943
SeAFusion	7.4713	18.0255	53.8434	61.9279	6.5330	0.5939	1.6560	0.4902
CrossFuse	7.1477	12.0273	43.9628	59.8653	4.4923	0.5231	1.1134	0.3541
ITFuse	6.3504	5.0112	24.7218	63.9461	2.1164	0.6373	1.0983	0.2087
DSFusion	7.0425	15.1334	38.6248	62.7078	5.8427	0.5426	1.1518	0.5474
CDDFuse	7.4038	17.1333	52.3959	61.7766	6.0774	0.6244	1.8486	0.4927
LEFuse	7.4624	16.5772	51.7002	61.8886	6.2857	0.6061	1.6579	0.4130
ST-RCAFuse	7.4936	18.0506	54.1449	62.8003	6.9798	0.6356	1.8505	0.4989

large areas of overexposure and significant suppression of background textures. Although the infrared image can clearly capture thermal targets such as vehicles and pedestrians, the discriminability of these targets may degrade under strong high-intensity illumination backgrounds. This dataset provides an effective benchmark for evaluating the performance of different fusion methods under extremely strong illumination interference conditions.

From the enlarged ROI regions in Fig. 9, it can be observed that under extreme conditions with strong headlight glare and traffic light illumination, existing comparison methods generally suffer from severe performance degradation. RFN-Nest, CrossFuse, and ITFuse are heavily affected by illumination overflow, where the headlight regions become excessively diffused, causing vehicle structures to be overwhelmed and rendering the targets unrecognizable. DenseFuse and NestFuse exhibit significant detail loss in the traffic light regions, with blurred and indistinct shapes. DSFusion introduces noticeable white noise in background areas such as the sky, degrading the natural appearance of the image. LEFuse can preserve the vehicle contours to some extent; however, the brightness of the headlights is suppressed. In contrast, the proposed ST-RCAFuse maintains stable fusion quality even under extremely strong illumination conditions. It not only preserves sharp and clear traffic light structures and recognizable vehicle shapes but also effectively enhances the headlight characteristics. A better balance is achieved among illumination suppression, target preservation, and detail fidelity, resulting in more natural and visually realistic fusion results.

Table 5 presents the objective evaluation results of ten fusion methods on the nighttime strong illumination interference dataset. It can be observed that the proposed ST-RCAFuse achieves the best performance in five metrics, including EN, SF, SD, AG, and SCD, while ranking third in CC and Qabf. By jointly considering both subjective visual quality and objective evaluation results, it is evident that the proposed ST-RCAFuse exhibits superior capability in information preservation, detail representation, and structural integrity under extremely strong illumination interference conditions. Overall, the fusion performance significantly outperforms the other comparison methods.

2.3.5 Analysis of fusion results on the ultra-low-light field dataset

Five representative pairs of source images are selected from the self-constructed ultra-low-light field dataset for fusion comparison, as shown in Fig. 10. All scenes correspond to complex field environments, including typical elements such as pedestrians, grasslands, and trees. In the first three groups, the visible images are severely degraded by extremely low illumination, where background structures are heavily corrupted by noise and fine details are difficult to discern. The last two groups simulate long-range detection scenarios, in which infrared targets occupy only a very small proportion of the image pixels. This dataset closely reflects real-world field monitoring conditions and provides an effective benchmark for evaluating the capability of different methods in preserving details and enhancing targets under extremely low-light conditions and small-target detection scenarios.

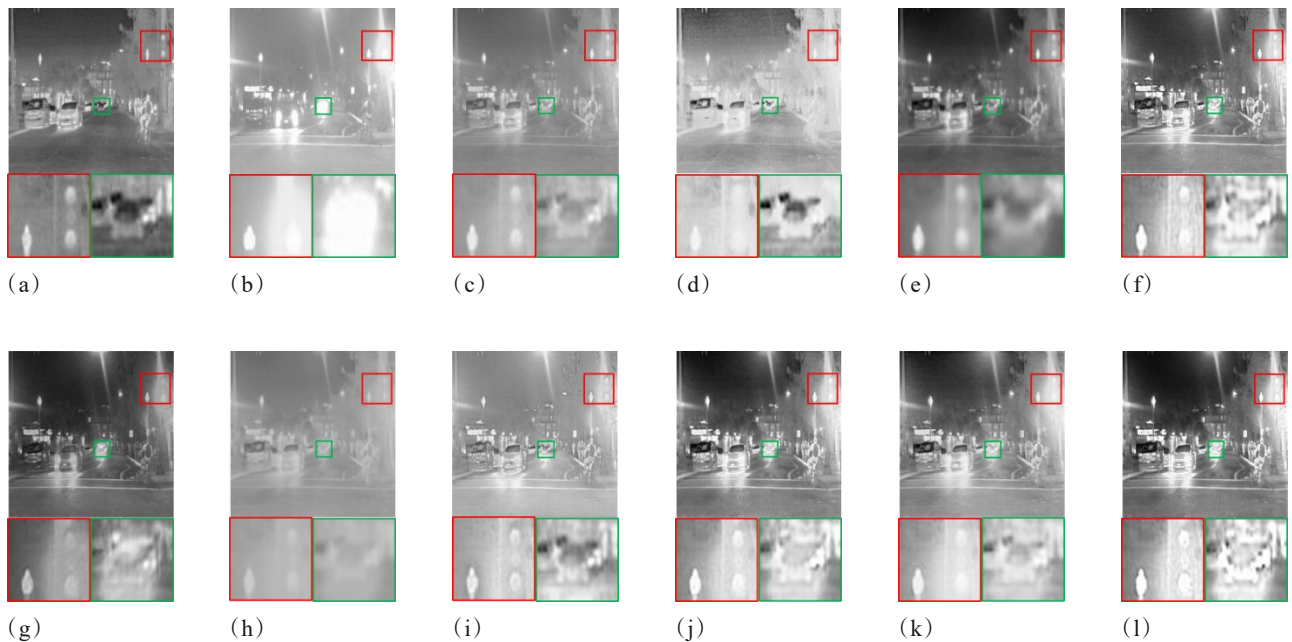
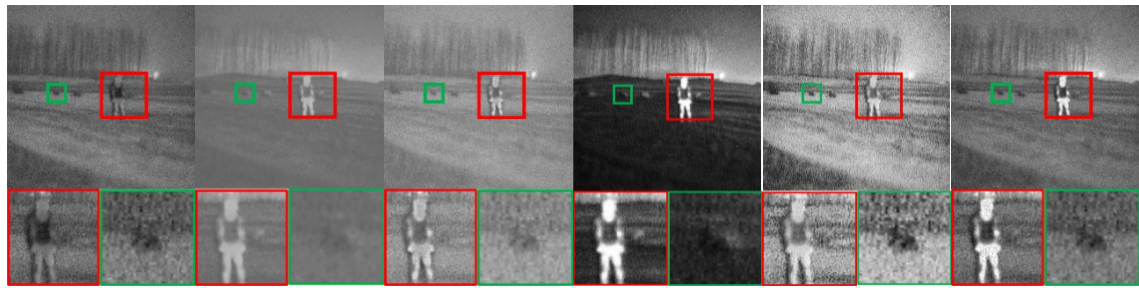
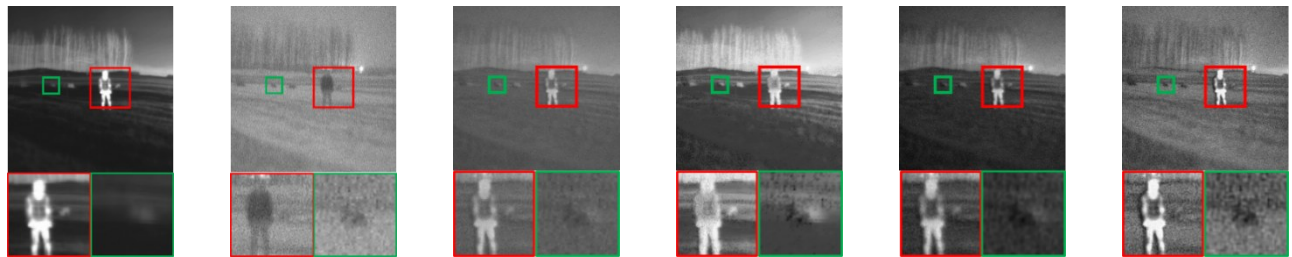


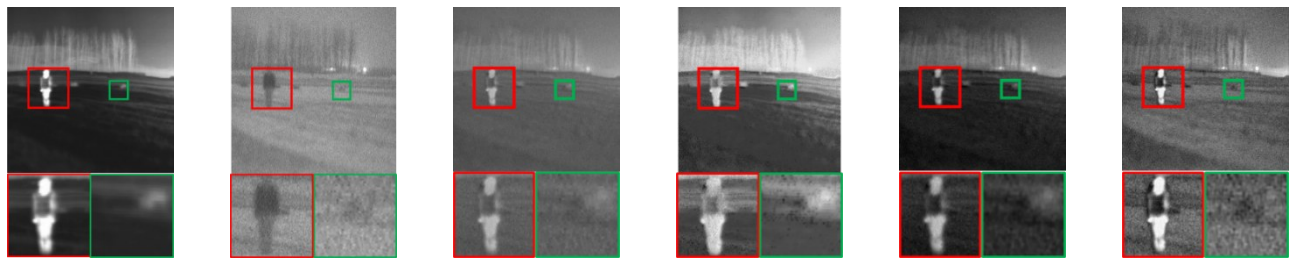
Fig. 9 Fusion results on the nighttime strong illumination interference test set: (a) Infrared image; (b) Visible image; (c) DenseFuse; (d) NestFuse; (e) RFN-Nest; (f) SeAFusion; (g) CrossFuse; (h) ITFuse; (i) DSFusion; (j) CDDFuse; (k) LEFuse; (l) Proposed ST-RCAFuse
图9 夜间强光干扰测试集上的融合结果: (a)红外图像; (b)可见光图像; (c) DenseFuse; (d) NestFuse; (e) RFN-Nest; (f) SeAFusion; (g) CrossFuse; (h) ITFuse; (i) DSFusion; (j) CDDFuse; (k) LEFuse; (l)本文 ST-RCAFuse

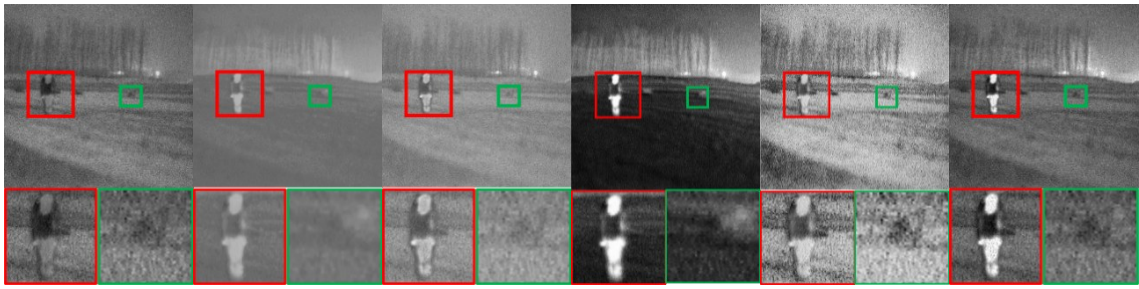
Table 5 Quantitative evaluation of ten fusion methods on the nighttime strong illumination interference test set
表5 夜间强光干扰测试集上10种融合方法的定量评价

	EN	SF	SD	PSNR	AG	CC	SCD	Qabf
DenseFuse	7.0584	8.1536	37.0219	66.2905	3.2199	0.8055	1.0818	0.4212
NestFuse	7.3761	13.0304	56.9765	62.1742	4.5699	0.7325	1.3494	0.3763
RfN-Nest	7.4107	6.8395	47.4827	62.5217	2.8629	0.8029	1.5541	0.2879
SeAFusion	7.5365	16.8902	58.0018	63.4523	5.6687	0.7589	1.6118	0.5305
CrossFuse	7.1244	10.4928	39.4564	61.3013	3.8491	0.7755	1.2651	0.3371
ITFuse	6.6577	4.9653	29.8827	65.5571	2.0859	0.7924	0.6395	0.2219
DSFuse	7.0577	14.1894	36.4475	65.0596	5.4459	0.7526	0.8768	0.5807
CDDFuse	7.5706	14.1867	56.6403	63.9824	5.6317	0.7948	1.6315	0.5528
LEFuse	7.5907	14.5623	57.3535	62.8151	5.5527	0.7668	1.5149	0.4142
ST-RCAFuse	7.6161	17.0782	58.2136	64.8296	6.0988	0.7967	1.6935	0.5342

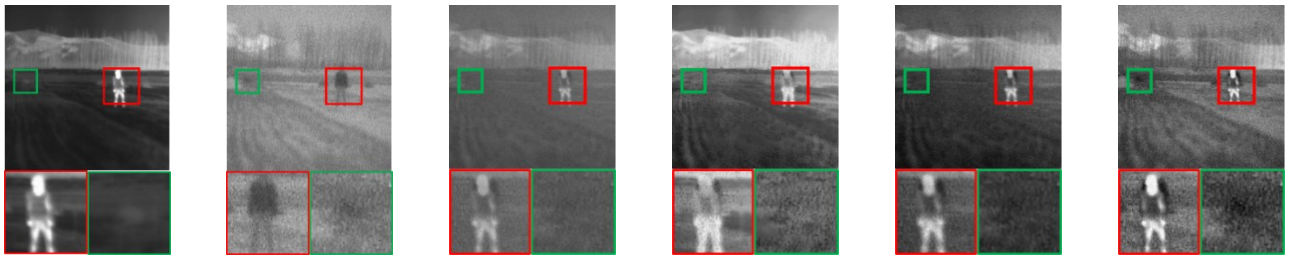


(a)

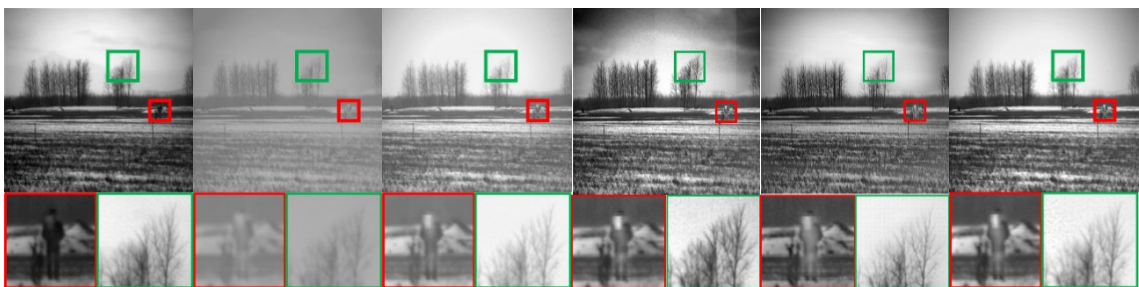
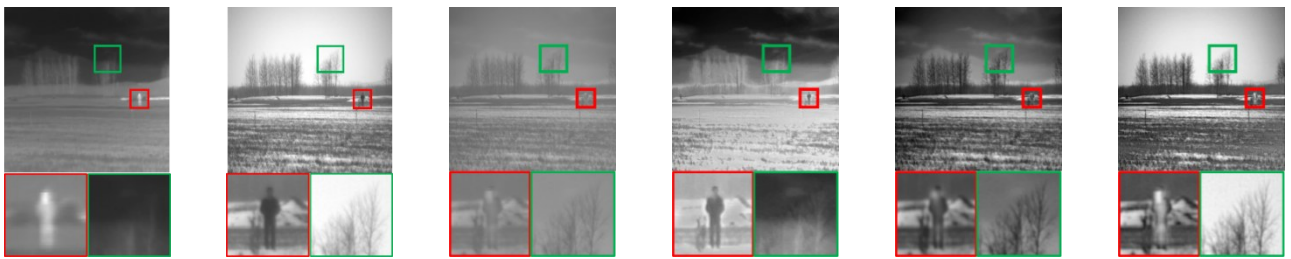




(b)



(c)



(d)

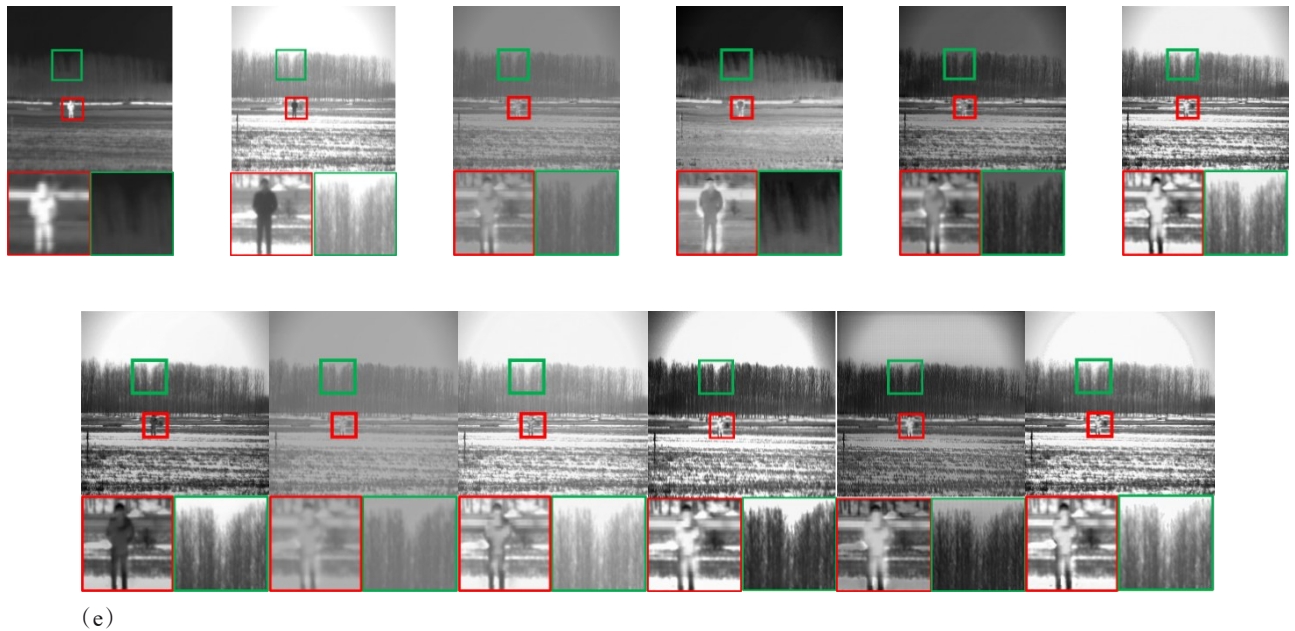


Fig. 10 Fusion results on the ultra-low-light field test set: (a)Scene 1; (b)Scene 2; (c)Scene 3; (d)Scene 4; (e) Scene 5. In each scene, the sub-images from left to right and top to bottom are: Infrared image, Visible image, fusion results of DenseFuse, NestFuse, RFN-Nest, SeAFusion, CrossFuse, ITFuse, DSFusion, CDDFuse, LEFuse, and the proposed ST-RCAFuse.

图 10 极低照度野外测试集上的融合结果: (a)场景 1; (b)场景 2; (c)场景 3; (d)场景 4; (e)场景 5。每组场景内,子图从左到右、从上到下依次为:红外图像、可见光图像、DenseFuse、NestFuse、RFN-Nest、SeAFusion、CrossFuse、ITFuse、DSFusion、CDDFuse、LEFuse 以及本文所提 ST-RCAFuse 的融合结果。

From the comparison results in Fig. 10, it can be observed that under extremely low-light conditions and weak-target scenarios in complex field environments, existing methods exhibit evident limitations. DenseFuse and ITFuse generate fusion images with low overall contrast, resulting in insufficient enhancement of infrared salient targets and blurred details. NestFuse suffers from severe feature attenuation, leading to the near-complete loss of fine textures, such as branches in Scenes 4 and 5. CrossFuse and DSFusion preserve certain background textures; however, their ability to enhance infrared targets is limited under low-light conditions, where the pedestrian targets in Scenes 1 - 3 appear dim and lack discriminability. RFN-Nest struggles to balance the large modality discrepancy under extreme conditions, causing the fusion results to be overly dominated by infrared information, with noticeably darkened sky regions in Scenes 4 and 5. CDDFuse is affected by low-light noise, producing evident linear vertical artifacts in the sky region of Scene 4, which disrupts spatial continuity. LEFuse, influenced by high-gain interference in extreme conditions, introduces significant granular noise in the background regions of Scenes 1 - 3. In contrast, the proposed ST-RCAFuse demonstrates strong robustness under these dual extreme challenges. It not only effectively enhances infrared salient targets but also achieves more accurate and complete restoration of fine details, leading to a superior balance between target enhancement and detail preservation.

Table 6 summarizes the average values of eight evaluation metrics for ten fusion methods on the ultra-low-light field dataset. The results show that the proposed ST-

RCAFuse ranks first in four key metrics, including SF, SD, AG, and Qabf, second in EN, and third in PSNR and CC. Overall, both subjective visual comparisons and objective quantitative results consistently demonstrate that ST-RCAFuse significantly outperforms existing methods, exhibiting stronger adaptability and robustness in extreme environments.

2.4 Ablation study

2.4.1 Ablation study on loss function weight coefficients

To balance the contributions of the feature decomposition loss L_{decomp}^I and the gradient consistency loss L_{grad}^I during Training Stage I, two weighting coefficients, α_1 and β_1 , are introduced for regulation. For the hyperparameter α_2 in Training Stage II, it is set as $\alpha_2 = \alpha_1$, since the feature decomposition loss in Stage II is identical to that in Stage I. Both parameters are used to constrain the feature decoupling process, and maintaining consistent weights ensures that the optimization criterion remains unchanged when transitioning from image reconstruction to the fusion task.

A systematic ablation study is conducted on the TNO dataset to evaluate the effects of different values of α_1 and β_1 , as shown in Table 7.

From the quantitative results on the TNO dataset, it can be observed that when $\alpha_1=2$ and $\beta_1=5$, the model achieves peak performance on several key metrics, including Entropy (EN), Spatial Frequency (SF), Standard Deviation (SD), Peak Signal-to-Noise Ratio (PSNR), and Sum of the Correlations of Differences (SCD). This indicates that this weight combination maximizes the information content and structural fidelity of

Table 6 Quantitative evaluation of ten fusion methods on the ultra-low-light field test set
表6 极低照度野外测试集上10种融合方法的定量评价

	EN	SF	SD	PSNR	AG	CC	SCD	Qabf
DenseFuse	6.1224	7.8224	19.5296	61.9016	2.9719	0.4799	1.6152	0.3883
NestFuse	7.3865	11.1492	39.0580	59.3081	3.7031	0.3585	1.1023	0.2866
RFN-Nest	7.0316	8.4083	37.7239	59.9116	3.9198	0.4603	1.9030	0.4943
SeAFusion	6.9535	18.0244	39.5604	60.6831	7.7375	0.4209	1.3144	0.6341
CrossFuse	6.8739	20.6111	37.6942	60.1580	7.6403	0.2994	0.5574	0.7227
ITFuse	5.7779	4.2513	17.1724	61.4311	1.8925	0.4510	1.4800	0.1822
DSFuse	6.3602	14.9804	27.3459	59.0963	6.0118	0.3336	1.0447	0.7544
CDDFuse	6.8908	17.1261	39.7420	58.7207	6.0933	0.4096	1.6061	0.5883
LEFuse	7.0548	19.0073	38.2508	57.9122	7.7662	0.3583	1.1487	0.3733
ST-RCAFuse	7.1202	21.9144	40.3424	60.7775	8.3516	0.4516	1.3522	0.7649

Table 7 Ablation study results of loss function weights α_1 and β_1 on the TNO dataset

表7 TNO数据集上损失函数权重系数 α_1 , β_1 消融实验结果

α_1	β_1	EN	SF	SD	PSNR	SCD
1	5	7.1386	14.0622	43.9597	62.4215	1.8028
2	5	7.1565	14.1998	45.4504	62.4230	1.8109
4	5	7.1348	14.0288	44.3212	62.3856	1.8055
2	2	7.1556	13.9116	45.4201	62.3476	1.8106
2	10	7.1456	14.1797	44.8038	62.4207	1.8085

the fused images. When α_1 is reduced to 1, the feature decomposition constraint becomes insufficient, making it difficult for the encoder to effectively distinguish between background and detail features. As a result, the saliency of fusion targets decreases, and all evaluation metrics show a significant decline. Conversely, when α_1 is increased to 4, the overly strong constraint leads to detail distortion and slight artifacts, causing performance degradation. For β_1 , when it is set to 2, the gradient constraint is relatively weak, resulting in insufficient edge and texture preservation in the reconstructed images, and the background details appear less prominent. When β_1 is increased to 10, the gradient constraint becomes excessively strong, leading to over-sharpened edges and degraded visual quality. The combination of $\alpha_1=2$ and $\beta_1=5$ achieves an optimal balance between feature decomposition constraint and gradient fidelity constraint. It not only ensures effective decoupling of background and detail features but also preserves clear edge and texture information.

2.4.2 Ablation study on the network architecture

To verify the functionality and effectiveness of each component module in the proposed ST-RCAFuse method, five different model configurations are constructed for ablation analysis on the TNO dataset. The details of the experimental settings are as follows, and the corresponding objective evaluation metrics calculated are shown in Table 8.

Experiment 1 (Baseline): A standard convolutional encoder-decoder network is used, trained only in the first stage.

Experiment 2 (Baseline + Two-stage): Builds on

Table 8 Ablation study results of the network architecture on the TNO dataset

表8 TNO数据集上网络架构的消融实验结果

	EN	SF	SD	PSNR	AG
Experiment 1	6.8868	11.4560	44.3535	60.5818	4.1422
Experiment 2	7.1324	12.4414	44.9263	61.4098	4.8780
Experiment 3	7.1458	12.8289	45.1211	61.5314	4.9653
Experiment 4	7.1489	13.8958	46.2506	61.3455	5.3395
Experiment 5	7.1565	14.1998	45.4504	61.4230	5.3518

Experiment 1 by incorporating the two-stage training strategy.

Experiment 3 (Baseline + Two-stage + RCAB): Builds on Experiment 2 by introducing the Residual Channel Attention Block (RCAB) as the detail extraction module in the encoder.

Experiment 4 (Baseline + Two-stage + RCAB + Swin Transformer): Further incorporates the Swin Transformer as the background modeling module.

Experiment 5 (Full ST-RCAFuse): The complete fusion network proposed in this work.

As shown in Table 8, compared with Experiment 1, Experiment 2, which incorporates the two-stage training strategy, demonstrates significant improvements across all objective evaluation metrics. This indicates that adding the fusion-stage training constraint enables the model to more effectively extract and integrate useful information from the source images. Building upon Experiment 2, Experiments 3 and 4 gradually introduce the RCAB and the Swin Transformer, respectively. Compared with Experiment 2, all objective metrics except PSNR show further improvement, demonstrating the positive contribution of these modules in enhancing feature representation and preserving fine details. Finally, the complete fusion model corresponding to Experiment 5 achieves the best overall performance across all evaluation metrics, validating the effectiveness and rationality of each component in the proposed ST-RCAFuse method.

The qualitative fusion results shown in Figure 11 further demonstrate the progressive and consistent improvements brought by the introduction of different modules. In Experiment 1, due to the relatively simple network ar-

chitecture and training strategy, the fused images appear overall blurry with low contrast and severe loss of texture information, for example, the branch textures in the red regions are blurred and difficult to distinguish; With the incorporation of the two-stage training strategy in Experiment 2, the overall contrast of the fused images is improved, and structural blurring is noticeably alleviated, allowing the branch textures in the red regions to be roughly recognizable; Building upon this, the addition of the RCAB module in Experiment 3 significantly enhances high-frequency details and edge structures, rendering textures such as the branches clearer and more complete, but the branch textures in the complex background of the green regions remain underrepresented; With the introduction of the Swin Transformer in Experiment 4, the model's ability to capture complex background structures is further strengthened, allowing the branch textures in the green regions to be effectively recovered, while overall structural consistency and background naturalness are improved. In contrast, Experiment 5, corresponding to the complete ST-RCAFuse model, achieves the best performance in terms of detail clarity, information representation in complex scenes, and overall visual quality, enabling a more complete and natural integration of key information from both infrared and visible images.

3 Conclusions

To address the limitations of insufficient background structure modeling and inadequate detail preservation in infrared and visible image fusion, a dual-branch image fusion network based on Swin Transformer Background Modeling and Residual Channel Attention Detail Enhancement (ST-RCAFuse) was proposed. In the encoding stage, the network employs a collaborative design of a background modeling branch and a detail enhancement branch to effectively capture global background structures and local high-frequency detail features. In the decoding stage, high-quality fused images are generated through progressive fusion and reconstruction. A multi-dimensional evaluation framework consisting of in-domain, conventional out-of-domain, and extreme scenarios is established. Extensive experiments are conducted

on the FLIR in-domain test set, the TNO and RoadScene conventional out-of-domain test sets, as well as two newly constructed extreme scenario datasets, including nighttime strong illumination interference and ultra-low-light field environments. The experimental results demonstrate that the proposed method effectively improves structural consistency and detail representation in fused images. It consistently achieves superior visual quality and objective performance not only in standard scenarios but also under challenging conditions such as strong illumination interference and extremely low-light environments, exhibiting enhanced robustness and generalization capability. Ablation studies further validate the effectiveness and complementarity of each component in enhancing fusion performance. Despite its robustness and generalization capability across various datasets and scenarios, ST-RCAFuse has certain limitations. Specifically, the current work primarily targets static image fusion and does not fully address dynamic scenes or real-time fusion requirements. Future work will focus on real-time fusion algorithms and applications in complex multi-modal scenarios, such as fast-moving targets, to further expand the practicality and application scope of the proposed method.

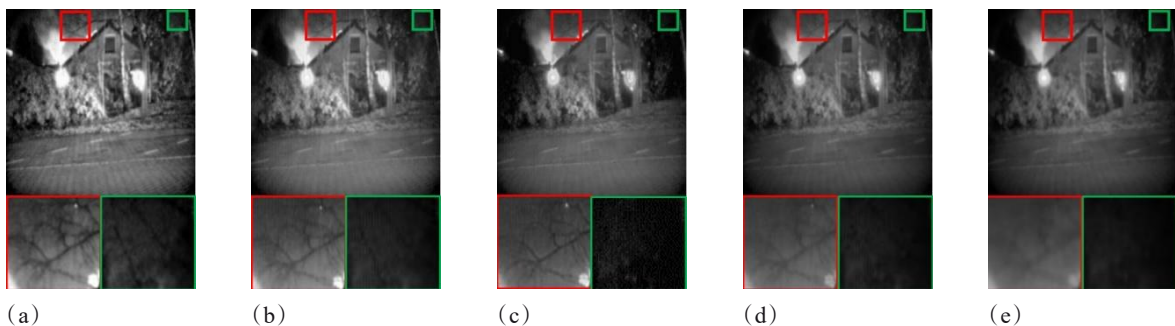


Fig. 11 Ablation study results of network architectures: (a) Experiment 1; (b) Experiment 2; (c) Experiment 3; (d) Experiment 4; (e) Experiment 5

图 11 网络架构的消融实验结果: (a)实验 1; (b)实验 2; (c)实验 3; (d)实验 4; (e)实验 5

References

- [1] Li S H, Cai W, Wang X, et al. A survey of infrared and visible image fusion methods based on a deep learning framework [J]. *Computer Engineering and Applications*, 2025, 61(9): 25-40.
(李淑慧,蔡伟,王鑫,等.深度学习框架下的红外与可见光图像融合方法综述[J].*计算机工程与应用*),2025,61(9):25-40.
- [2] Chen F, Wang N, Tang J, et al. Unsupervised person re-identification via multi-domain joint learning[J]. *Pattern Recognition*, 2023, 138: 109369.
- [3] Bai L, Zhang W, Pan X, et al. Underwater image enhancement based on global and local equalization of histogram and dual-image multi-scale fusion [J]. *IEEE Access*, 2020, 8: 128973-128990.
- [4] Shen Y. RGBT bimodal twin tracking network based on feature fusion[J]. *Journal of Infrared and Millimeter Waves*, 2021, 50(3):20200459.
- [5] Cong T, Yongshun L, Hua Y, et al. Decision-level fusion detection for infrared and visible spectra based on deep learning [J]. *Infrared and Laser Engineering*, 2019, 48(6): 626001.
- [6] Li S, Yang B, Hu J. Performance comparison of different multi-resolution transforms for image fusion [J]. *Information fusion*, 2011, 12(2): 74-84.
- [7] Wang Z, Xu J, Jiang X, et al. Infrared and visible image fusion via hybrid decomposition of NSCT and morphological sequential toggle operator[J]. *Optik*, 2020, 201: 163497.
- [8] Zhang X, Ma Y, Fan F, et al. Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition [J]. *Journal of the Optical Society of America A*, 2017, 34(8): 1400-1410.
- [9] Fu Z, Wang X, Xu J, et al. Infrared and visible images fusion based on RPCA and NSCT [J]. *Infrared Physics & Technology*, 2016, 77: 114-123.
- [10] Zong J, Qiu T. Medical image fusion based on sparse representation of classified image patches[J]. *Biomedical Signal Processing and Control*, 2017, 34: 195-205.
- [11] 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.
- [12] Li H, Wu X J, Durrani T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. *IEEE Transactions on Instrumentation and Measurement*, 2020, 69(12): 9645-9656.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] Li H, Wu X J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach[J]. *Information Fusion*, 2024, 103: 102147.
- [17] Tang W, He F, Liu Y. ITFuse: An interactive transformer for infrared and visible image fusion [J]. *Pattern Recognition*, 2024, 156: 110822.
- [18] Cheng M, Huang H, Liu X, et al. LEFuse: Joint low-light enhancement and image fusion for nighttime infrared and visible images[J]. *Neurocomputing*, 2025, 626: 129592.
- [19] Song W, Li Q, Gao M, et al. SFINet: A semantic feature interactive learning network for full-time infrared and visible image fusion [J]. *Expert Systems with Applications*, 2025, 261: 125472.
- [20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. *Advances in neural information processing systems*, 2017, 30:5998-6008
- [21] Wang Z, Chen Y, Shao W, et al. SwinFuse: A residual swin transformer fusion network for infrared and visible images [J]. *IEEE Transactions on Instrumentation and Measurement*, 2022, 71: 1-12.
- [22] Toet A, Hogervorst M A. Progress in color night vision [J]. *Optical Engineering*, 2012, 51(1): 010901.
- [23] 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.
- [24] Tang L, Deng Y, Ma Y, et al. SuperFusion: A versatile image registration and fusion network with semantic awareness [J]. *IEEE/CAA Journal of Automatica Sinica*, 2022, 9(12): 2121-2137.
- [25] Roberts J W, Van Aardt J A, Ahmed F B. Assessment of image fusion procedures using entropy, image quality, and multi-spectral classification [J]. *Journal of Applied Remote Sensing*, 2008, 2(1): 023522.
- [26] Eskicioglu A M, Fisher P S. Image quality measures and their performance[J]. *IEEE Transactions on communications*, 2002, 43(12): 2959-2965.
- [27] Rao Y J. In-fibre Bragg grating sensors [J]. *Measurement science and technology*, 1997, 8(4): 355.
- [28] Jagalingam P, Hegde A V. A review of quality metrics for fused image[J]. *Aquatic Procedia*, 2015, 4: 133-142.
- [29] 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.
- [30] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. *Electronics letters*, 2000, 36(4): 308-309.
- [31] 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.
- [32] Tang L F, Zhang H, Xu H, et al. A survey of image fusion methods based on deep learning [J]. *Journal of Image and Graphics*, 2023, 28(1): 3-36.
(唐霖峰,张浩,徐涵,等.基于深度学习的图像融合方法综述[J].*中国图象图形学报*),2023,28(1):3-36.
- [33] Liu K, Li M, Chen C, et al. DSFuse: Infrared and visible image fusion method combining detail and scene information [J]. *Pattern Recognition*, 2024, 154: 110633. Innovations of this Work:
- [34] A dual-branch collaborative infrared and visible image fusion network is proposed, which decouples and models low-frequency background structural information and high-frequency texture detail information through a background branch and a detail branch. Specifically, the background branch combines coordinate attention and Swin Transformer to enhance directionality in the feature space and model both global and local background structures; The detail branch employs a residual channel attention module to strengthen the representation of texture details and edge structures, thereby improving the structure retention capability and detail expressiveness of the fused image.
- [35] A two-stage training strategy is proposed, consisting of feature

decomposition-reconstruction and feature fusion-generation, along with a multi-constraint composite loss function. In the first stage, reconstruction constraints ensure effective feature decomposition and faithful image reconstruction. In the second stage, joint constraints on pixel information and structural information enhance the information retention and structural consistency of the fused images, thereby improving the fusion performance and

robustness of the model.

[36] An ultra-low-light field test dataset is constructed, and comparative as well as generalization experiments are conducted across multiple datasets. The results demonstrate that the proposed method achieves strong fusion performance in complex scenarios such as low-light environments, thereby extending the application scope of infrared and visible image fusion methods.

基于 Transformer 背景建模与 CAM 细节增强的红外与可见图像融合

纪肖剑¹, 喻春雨¹, 陈陆杰¹, 张俊举², 孙斌³

(1. 南京邮电大学 电子与光学工程学院、柔性电子(未来技术)学院, 江苏 南京 210023;

2. 南京理工大学 电子工程与光电技术学院, 江苏 南京 210094;

3. 南京邮电大学 自动化学院, 江苏 南京 210023)

摘要: 针对红外与可见光图像融合中背景结构建模不足、细节纹理表达不充分的问题, 提出一种基于 Swin Transformer 背景建模与残差通道注意力细节增强的双分支图像融合网络 (Swin Transformer Background Modeling and Residual Channel Attention Detail Enhancement for Image Fusion, ST-RCAFuse), 具体方法是: 在编码阶段, 设计背景与细节两个分支。其中, 背景分支以 Swin Transformer 为核心, 通过窗口自注意力机制 (Window-based Self-Attention, WSA) 实现全局与局部背景结构的高效建模, 并引入坐标注意力机制 (Coordinate Attention, CoordAtt) 以增强特征的空间方向性; 细节分支采用残差通道注意力模块 (Residual Channel Attention Block, RCAB) 以提取纹理细节与高频信息。在解码阶段, 背景和细节两类特征通过逐级融合与重建生成高质量融合图像。实验选择 FLIR 数据集进行网络训练, 并构建了“域内+常规域外+极端场景”的多维度测试体系; 选取 FLIR 测试集作为域内测试组验证同源数据分布下的基础性能, 选取 TNO、RoadScene 数据集作为常规域外测试组验证跨场景泛化能力, 同时采用夜间强光干扰测试集与极低照度野外测试集, 全面评估模型在复杂恶劣环境下的鲁棒性。实验结果表明, 所提 ST-RCAFuse 在常规公开数据集上融合视觉效果最优, 信息熵、空间频率、标准差、平均梯度等核心指标均取得领先; 在夜间强光干扰、极低照度野外等极端场景下, 仍能有效抑制干扰、保留细节和增强目标, 融合性能显著优于现有对比方法, 充分验证了其在多场景、多极端条件下的优异泛化能力、鲁棒性与实用价值。

关键词: 双分支; 图像融合; 双阶段训练; 坐标注意力; 残差通道注意力

中图分类号: TP391

文献标识码: A