Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AdaFuse: a cross-modal image fusion algorithm based on negative value calibration and adaptive hierarchical activation

Yongxin Wang

wangyongxin@hrbust.edu.cn

Harbin University of Science and Technology

Yan Yang

Harbin University of Science and Technology

Jiaqi Yuan

Harbin University of Science and Technology

Qinghua Liu

Harbin University of Science and Technology

Baoyuan Chen

Harbin University of Science and Technology

Research Article

Keywords: Precise Segmentation, Image Fusion, Adaptive Activation, Negative Value Calibration

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10828474/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AdaFuse: a cross-modal image fusion algorithm based on negative value calibration and adaptive hierarchical activation.

Yongxin Wang *, Yan Yang, Jiaqi Yuan, Qinghua Liu, Baoyuan Chen

School of measurement and communication engineering, Harbin University of Science and Technology, Harbin 150080, China

*Correspondence:Yongxin Wang: wangyongxin@hrbust.edu.cn

Abstract：For cross-modal image fusion, fused images often suffer from imbalanced contributions of cross-modal features, while multi-channel activation results in a significant increase in intermediate tensor dimensions within the fusion network. To address these challenges, we propose AdaFuse, a cross-modal image fusion algorithm based on negative value calibration and adaptive hierarchical activation. By employing a hierarchical feature perception strategy and dynamic computational allocation, the algorithm routes feature maps into shallow or deep processing paths based on their information content, enabling adaptive allocation of computational resources according to information density. During the fusion stage, negative value calibration is used to select attention regions for modal guidance, suppressing conflicting activations that may cause image distortion. We further introduce CA-Net, a cross-modal discriminator that incorporates multi-pathway modal information to enhance feature representation during training. Visible-infrared image fusion is selected as the validation task for the proposed algorithm. Extensive experiments on widely adopted benchmarks demonstrate that our approach significantly outperforms existing methods in terms of computational efficiency and fusion quality. Moreover, the proposed framework of hierarchical perception and precise control provides a promising methodological reference for broader tasks, such as hyperspectral and medical image fusion.

Key Words: Precise Segmentation; Image Fusion; Adaptive Activation; Negative Value Calibration

1 INTRODUCTION

Among the various research directions in multi-source image fusion, infrared and visible image fusion has attracted considerable attention in both academia and engineering due to the complementary characteristics of the two modalities in complex environments[1]. Infrared images capture thermal radiation and enable robust target detection under low-illumination conditions, such as nighttime and adverse weather, whereas visible images preserve rich texture, structural, and color details that are crucial for human perception and subsequent vision tasks. Effective fusion of these modalities not only enhances target detectability but also improves the overall visual quality of the fused image, thereby contributing to advanced intelligent perception systems [2].

Traditional fusion methods—such as those based on multi-scale transforms (e.g., wavelet transforms[3] and sparse representation [4], can be viewed as saliency-driven weighting strategies. While they have achieved moderate success, they are inherently limited in handling complex nonlinear relationships and in extracting high-level semantic information. With the advent of deep learning, fusion approaches leveraging Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) have become mainstream[5]. These methods learn fusion strategies directly from data in an end-to-end manner and have demonstrated clear advantages over conventional approaches.

Liu et al. [6] designed a residual fusion module based on ResNet to fuse residual information rather than the original source images. This strategy helps preserve visible color and texture while introducing complementary infrared information, thereby reducing color distortion. However, the residual design introduces a large number of intermediate feature maps, increasing channel concatenation complexity and computational cost. Zhao et al.[7] employed deep image decomposition for region-aware fusion by separating the image into a base layer and a thermal-detail layer. This design enables finer control of infrared information injection and alleviates oversaturation caused by excessive thermal energy in the background. However, its strong dependence on segmentation-oriented optimization limits generalization across datasets, and noise in complex scenes may also be amplified when treated as detail information. Xu et al. [8] proposed an adaptive fusion loss function based on information preservation degree. This function dynamically adjusts the fusion weights according to the information content of the input image pair, thereby preventing one modality (especially the high-energy infrared image) from overly dominating the fusion process and fundamentally alleviating oversaturation. Nevertheless, its strong pursuit of generalization results in performance on final fusion metrics that is inferior to specifically constrained models, while the complex network layers also lead to lower computational efficiency compared to deep networks specifically designed for fusion.

To address these limitations, this paper proposes an adaptive hierarchical activation network for infrared-visible image fusion. The proposed framework introduces negative value calibration to regulate the diffusion of infrared features, employs a center-guided diffusion attention mechanism, and incorporates multiple perception-driven loss functions together with a comprehensive evaluation system. As a result, the method effectively improves both structural clarity and visual quality in the fused images.

2 PROPOSED ALGORITHM

2.1 Adaptively Activated Fusion for Visible and Infrared Images

In this paper, we propose a network architecture based on adaptive hierarchical activation connections. This architecture adaptively stratifies the visible image, directing layers with high mean deviation into the deep network for high-level semantic feature extraction, while the remaining layers are processed by the shallow network to preserve foundational details. The output from the shallow network serves as connecting layers, which are ultimately aligned with the deeply processed high-mean layers via feature alignment. By connecting the original image feature maps with intermediate processing layers, details are retained while effectively preventing the introduction of excessive tensors in the middle stages of the network, thereby significantly controlling computational complexity. Furthermore, this paper introduces a negative value calibration method with restricted diffusion. This method employs a gradient variation adjustment strategy between network layers, enabling the diffusion gradient to synchronize with the degree of variation in each layer. In summary, the proposed network can substantially reduce computational load compared to traditional methods while ensuring that the final fused image achieves enhanced thermal target representation and preserved background details.

2.1.1 Adaptive Convolutional Network Based on Visible Image Hierarchical Activation

The activation mechanism is a fundamental component of deep neural networks, as it introduces nonlinear transformations that enable the model to approximate complex nonlinear mappings. Rectified Linear Unit (ReLU) and its variants, such as Leaky ReLU, ELU, and GELU, have been widely adopted in convolutional neural networks (CNNs) and generative adversarial networks (GANs) because of their effectiveness in sparse activation and gradient propagation. To improve the quality of infrared-visible fused images while maintaining a reasonable computational cost, we introduce channel attention for adaptive feature activation, aiming to balance global contrast enhancement and local structure preservation.

Channel attention mechanisms, represented by the Squeeze-and-Excitation (SE) network [9], have substantially improved feature representation in deep models. By modeling inter-channel dependencies, they allow the network to adaptively recalibrate channel-wise responses, thereby strengthening informative features while suppressing redundant ones.

In low-light visible images, the RGB channels often exhibit noticeable differences in mean intensity, meaning that one channel may contain a larger proportion of dominant grayscale information. During forward propagation, this high-response channel tends to preserve more complete feature details because of its stronger signal intensity. In contrast, subtle features in low-response channels may be compressed into a narrower dynamic range after operations such as batch normalization, since they are jointly normalized with high-value channels. This distribution shift makes weak features more likely to fall into the inactive

region of subsequent activation functions, such as ReLU, thereby limiting their effective propagation and gradually weakening their representation in deeper layers.

To mitigate the issue of gradient diminishment or even vanishing in color domain layers with low mean values after processing through deep networks, this paper proposes a detail-first adaptive hierarchical activation fusion network. The specific pipeline is illustrated in Figure 1. It begins by performing global feature modeling separately on each RGB layer of the visible (VIS) image. Subsequently, a cross-modal cross-attention mechanism (LowEnhance & IR Guide) is utilized to achieve modality guidance and deep information interaction. As detailed in the layer stitching process shown in Figure 2, the single-channel fusion layer obtained from shallow network processing serves as a residual connection. This design can replace traditional ResNet layers, thereby simplifying channel concatenation operations within the network.

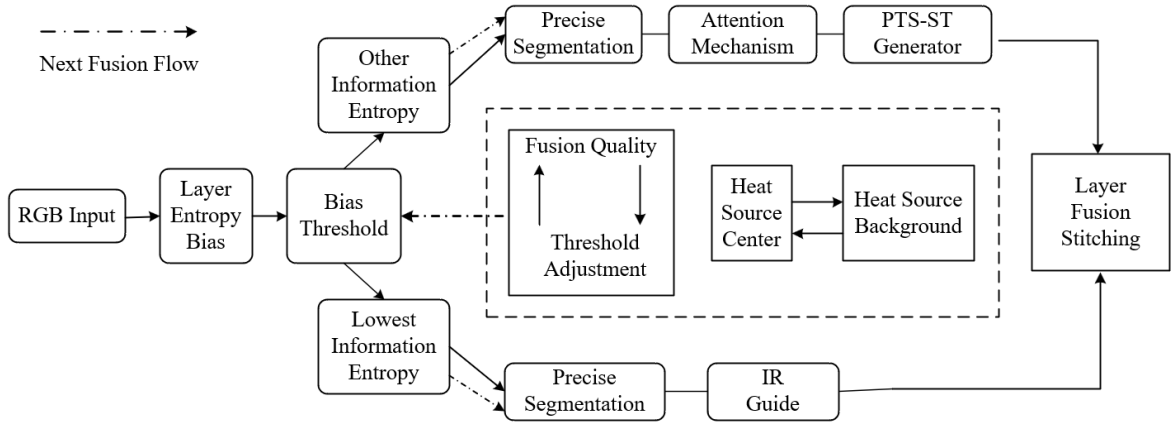


Fig.1. Adaptive Hierarchical Activation Process

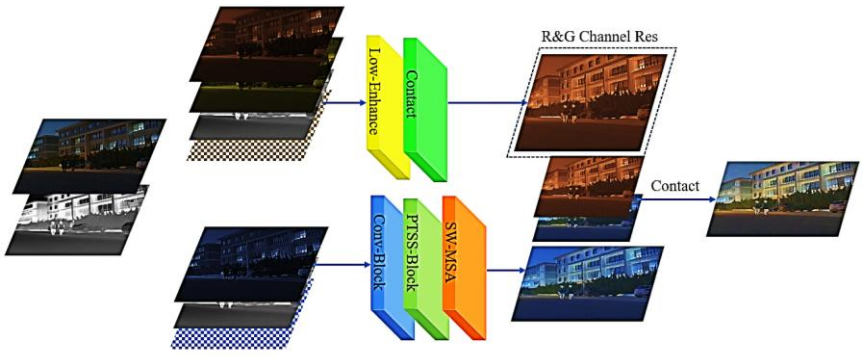


Fig.2. Dual-Stream Fusion

Linear projections are applied to the feature tensors of the infrared-visible image pair. Specifically, separate linear mappings are performed on the infrared features and on each RGB channel of the visible features, as expressed in Equation (1):

$$Q = W_q F_{IR} \in R^{d\times N}$$
$$K = W_k F_{iVIS} \in R^{d\times N} \quad (1)$$
$$V = W_v F_{iVIS} \in R^{d\times N}$$

Where F_{iVIS} represents the i-th color channel of the visible image, with $i \in [R, G, B]$,and $W_q, W_k, W_v \in R^{d\times C}$, are the learnable projection matrices. Here, d denotes the transformed feature dimension and $N$is the number of spatial tokens.

The attention weights are computed from the correlation between the Query and Key representations. In our design, the Value branch is emphasized to preserve fine details and maintain accurate texture contrast in the fused image. Channels with large Value-deviation ratios are routed to the backbone for deep processing. After concatenation with infrared features, negative value calibration with restricted diffusion is applied to suppress the excessive spread of enhanced infrared responses, thereby reducing color deviation and visual distortion in the fused result.

In contrast, channels with small Value deviations are processed in the shallow branch. At this stage, an infrared-guided low-frequency enhancement module (LowEnhance) is introduced to establish cross-modal interaction and enhance thermal target contours in the background. Through this dual-path parallel design, the overall computational burden is reduced because the shallow branch is significantly less expensive than the deep branch. In nighttime scenes with severe color cast, the computational cost can be further decreased because fewer channels are activated for deep processing.

2.1.2 Precise Segmentation of Infrared Thermal Targets

In conventional visible and infrared image fusion networks, the energy from high-temperature regions in the infrared image often diffuses indiscriminately during the fusion process [10]. This diffusion causes abnormal brightening in the detailed regions surrounding target contours, a phenomenon referred to as highlight contamination. As a result, target saliency is weakened and the fidelity of non-thermal background details is degraded.

To address this issue, this paper proposes a limited diffusion method for thermal targets based on spatial range constraints and negative value symmetric calibration. The core objectives of this method are to: enhance the saliency of thermal target regions while effectively preserving

surrounding environmental details, and prevent the disordered diffusion of high-temperature energy to adjacent pixels. First, a threshold segmentation is applied to the infrared image F_{IR} to extract the thermal target mask M_{IR} as shown in Equation (2):

$$M_{IR}(h,w) = \begin{cases} 1, & F_{IR}(h,w) \geq T_{IR} \\ 0, & otherwise \end{cases} \quad (2)$$

where, T_{IR} denotes the segmentation threshold, and (h,w) represents the spatial coordinates of each pixel. $M_{IR}(h,w) = 1$ indicates a high-temperature region, whereas $M_{IR}(h,w) = 0$ corresponds to the background. During fusion, enhancement of thermal responses is restricted to the spatial region defined by $M_{IR} = 1$. For each high-temperature pixel (h_0, w_0), an $s \times s$ window is centered on it. Corresponding feature patches are then extracted from the thermal region and from a symmetric non-thermal background region, denoted as the thermal window $\mathcal{W}_{hot}$ and the background window $\mathcal{W}_{bg}$, respectively. The thermal window centered at (h_0, w_0) is defined as

$$\mathcal{W}_{hot} = F_{IR}[h_0 - \Delta : h_0 + \Delta, w_0 - \Delta : w_0 + \Delta]$$

$$\mathcal{W}_{bg} = F_{VIS}[h_0 - \Delta : h_0 + \Delta, w_0 - \Delta : w_0 + \Delta] \quad (3)$$

$$\Delta = \min\left(\Delta_{\max}, \left\lfloor \frac{F_{IR}^{\max}}{\beta(h,w)} \right\rfloor\right)$$

Where Δ determines the spatial radius of the local window. The background window $\mathcal{W}_{bg}$ is selected from a symmetric non-thermal region with respect to the thermal center. For each pixel inside $\mathcal{W}_{hot}$, negative value calibration is imposed on pixels in the surrounding non-thermal region at an equivalent spatial distance. The key idea is to apply a calibration window of the same size around high-temperature pixels while using spatially symmetric background points as constraints.

During the later upsampling stages, this strategy creates a clearer gradient contrast between thermal and non-thermal regions, thereby enhancing thermal target saliency while suppressing excessive diffusion to nearby non-target pixels. In the fusion network, saliency detection or threshold segmentation is first performed on the infrared image to obtain a thermal saliency mask, M_{IR}.Under conventional attention mechanisms—such as Self-Attention (SA)[11], Co-Attention (CA)[12], and Window-Attention (WA)[13], high-response thermal features are further emphasized after attention computation. However, after downsampling and subsequent upsampling, these strong responses tend to spread into neighboring low-temperature regions, producing pseudo-halo artifacts.

To address this issue, this paper introduces a Distance-Decay Modulation with Suppressed Diffusion. This mechanism achieves a gradient attenuation, radiating outward from the central pixel where negative value calibration is applied, as detailed in Equation (4), The distance-decay coefficient is defined as

$$\beta(h,w) = max\left(exp\left(-\frac{d^2(h,w)}{2\sigma^2}\right), \beta_{bg}\right) \quad (4)$$

where $d(h,w)$ denotes the distance from the pixel at location(h,w) to the nearest thermal target boundary. The parameter σ controls the rate of attenuation, and β_{bg} denotes the minimum brightness gain for the background region. A learnable strategy is adopted for β_{bg} to adapt to images with varying contrast levels.

A visualization of this diffusion mechanism is presented in Figure 3, and its concrete implementation is given by Equation (5):

$$F_{cal} = \begin{cases} F_{IR}(h,w), & (h,w) \in \mathcal{W}_{hot} \\ 1 - \beta \cdot F_{IR}(h_s,w_s), & (h_s,w_s) \in \mathcal{W}_{bg} \end{cases} \quad (5)$$

where(h_s, w_s) denotes the background pixel symmetrically corresponding to the thermal center (h_0, w_0) and , $\beta > 0$ is the suppression coefficient defined in Equation (4). Here, F_{IR} is normalized to [0,1] . Therefore, the term $1 - \beta(h,w)F_{IR}(h_s,w_s)$ serves as a symmetric suppression response for non-thermal background regions. This coefficient is determined by the local distance-decay pattern around the thermal target. The calibrated feature map F_{cal} is then fed into the upsampling module to recover the original spatial resolution:

$$F_{up}(h,w) = Upsample\left(F_{cal}(h,w)\right) \quad (6)$$

Because thermal regions are positively preserved while non-thermal regions are suppressively calibrated, a clear gradient contrast is formed at the boundary between thermal and non-thermal areas. This effect can be measured by

$$G_{edge}(h,w) = \left|F_{up}(h,w) - F_{up}(h',w')\right| \quad (7)$$

where $(h',w') \in \mathcal{N}(h,w)$, $\mathcal{N}(h,w)$ denotes the non-thermal neighborhood of pixel (h,w). This mechanism can be explicitly integrated into the feature flow of the fusion network and further strengthened by the edge loss term:

$$L_{edge} = \frac{1}{N}\sum_{h,w} G_{edge}(h,w) \quad (8)$$

where N denotes this is used in the calculation of the loss function. σ denotes a non-linear activation function, and F_{VIS} represents the visible light features.

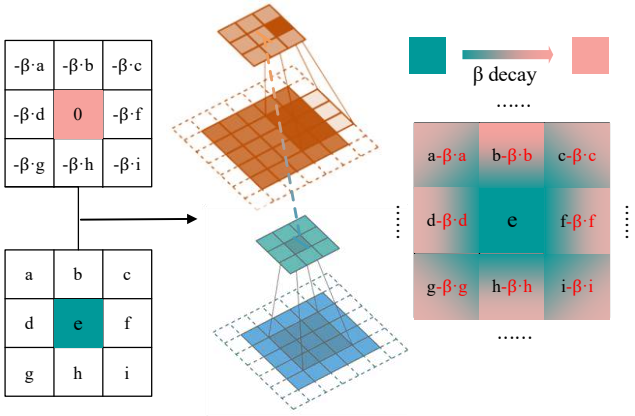


Fig.3. Segmentation algorithm

2.2 Loss Function Design

The proposed method mainly consists of an adaptive cross-activation architecture and a precise thermal-source segmentation module. Through specially designed loss functions, the model is encouraged to fully exploit the ability of DDCGAN to balance multi-source information preservation. As a result, the fused images not only preserve critical information from the infrared and visible inputs, but also better satisfy the requirements of human visual perception and downstream vision tasks.

2.2.1 Adaptive Activation Loss

To preserve low-detail and low-brightness features that may otherwise be weakened in excessively deep networks, and to adaptively determine whether visible features should be routed to deeper processing, we introduce the adaptive activation loss L_{AAM}, defined as

$$L_{\text{AAM}} = \frac{1}{N}\sum_{h,w}\left[\alpha(h,w) \cdot |F_{\text{out}}(h,w) - F_{IR}(h,w)| + \left(1 - \alpha(h,w)\right) \cdot |F_{\text{out}}(h,w) - F_{VIS}(h,w)|\right] \quad (9)$$

where $F_{out}(h,w)$denotes the fused output value (or feature response) at position (h,w) , while $F_{IR}(h,w)$ and $F_{VIS}(h,w)$ denote the corresponding infrared and visible values (or features), respectively. The adaptive weight $\alpha(h,w) \in [0,1]$is dynamically generated by the Adaptive Activation Module (AAM) according to local content. It indicates the relative preference of the fused output toward the infrared or visible modality: a larger $\alpha(h,w)$ places greater emphasis on infrared information, whereas a smaller value favors visible features. In practice, $\alpha(h,w)$can be jointly determined by the thermal mask, regional saliency, feature intensity, and learnable network parameters.The L_{AAM} loss primarily consists of two terms:
1.Infrared Guidance Term: $\alpha(h,w) \cdot |F_{out}(h,w) - F_{IR}(h,w)|$. In regions with a large $\alpha(h,w)$, the network output is strongly constrained to resemble the infrared features, reflecting infrared dominance (e.g., in thermal target regions).
2.Visible Guidance Term: $[1-\alpha(h,w)] \cdot |F_{\mathrm{out}}(h,w) - F_{VIS}(h,w)|$. In regions with a small $\alpha(h,w)$, the output is primarily guided to match the visible features, emphasizing visible light dominance (e.g., in environmental, detail-rich areas). Here, in $\frac{1}{N}\sum_{h,w}$ the N is the total number of pixels in the image, and the final loss is the average over all pixels.

2.2.2 Precise Thermal-Source Segmentation Loss

To encourage strong responses in thermal-source regions $(M_{IR}(h,w)=1)$ while suppressing direct thermal contamination in nearby non-thermal areas, this paper introduces a center- and neighborhood-weighted segmentation loss. This loss enhances thermal target saliency, allows moderate luminance transition in adjacent regions, and preserves structural details in distant background areas without introducing abnormal enhancement or excessive darkening. It is defined as

$$L_{mod} = \lambda_c \cdot L_c + \lambda_n \cdot L_n + \lambda_\beta \cdot L_\beta \quad (10)$$

where L_c(Eq. 11), L_n(Eq. 13), and L_β(Eq. 14) denote the thermal-center constraint, thermal-neighborhood constraint, and background constraint, respectively, while λ_c, λ_n, and λ_b are their corresponding weighting coefficients. The thermal-center constraint is formulated as

$$L_{\mathrm{c}} = \frac{1}{|M_{IR}|}\sum_{(h,w)\in M_{IR}} ||F_{\mathrm{out}}(h,w) - F_{IR}(h,w)|| \quad (11)$$

and the corresponding fused response is computed as

$$F_{out}(h,w) = \beta(h,w)F_{IR}(h,w) + [1-\beta(h,w)]F_{VIS}(h,w) \quad (12)$$

Here, $L_c$encourages the output in thermal-center regions to remain strongly aligned with infrared responses while still retaining visible contextual information. This design prevents uncontrolled diffusion of infrared energy and allows a moderate luminance elevation around the thermal source. In Eq. (12), $\beta(h,w)$is the calibration coefficient produced by the precise segmentation mechanism.The neighborhood constraint is defined as

$$L_n = \frac{1}{|N_{IR}|}\sum_{(h,w)\in N_{IR}} ||F_{\mathrm{out}}(h,w) - \left[\beta(h,w)F_{IR}(h,w) + \left(1-\beta(h,w)\right)F_{VIS}(h,w)\right]|| \quad (13)$$

where N_{IR}denotes the set of pixels surrounding the thermal center region. This term complements the center loss by constraining the local transition structure around thermal targets, thereby helping the model generate more reasonable boundaries, richer textures, and clearer details.To further regulate distant background regions, a background constraint term is introduced:

$$L_B = \frac{1}{|B|}\sum_{(h,w)\in B} ||F_{out}(h,w) - \left[\beta_{bg}F_{IR}(h,w) + \left(1-\beta_{bg}\right)F_{VIS}(h,w)\right]|| \quad (14)$$

where Bdenotes the background region, and $F_{IR}(h,w)$and $F_{VIS}(h,w)$ are the infrared and visible inputs at location (h,w), respectively. In thermal-center regions, $\beta(h,w) \approx 1$, so the output is dominated by infrared information and the targets remain salient. In neighborhood regions, $\beta(h,w)$ decreases gradually, allowing adaptive luminance transition while preventing direct contamination of nearby details. In distant background regions, $\beta(h,w)$is reduced to β_{bg}, so the output is mainly determined by visible features with only limited infrared contribution, which prevents the background from becoming overly dark without causing over-enhancement or unnatural contrast.

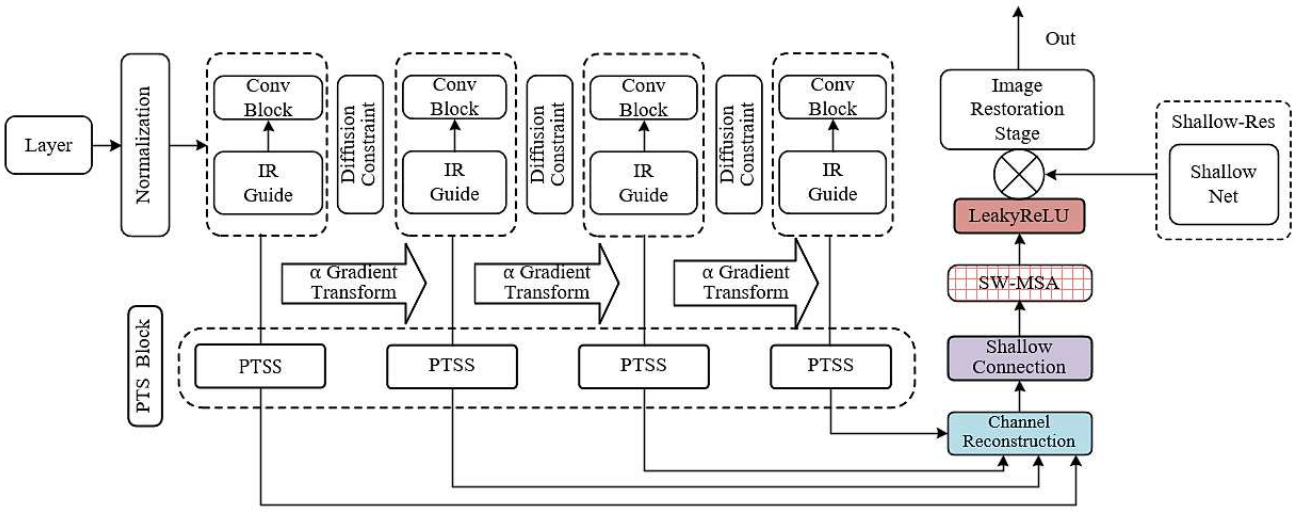


Fig.4. PTS Generator

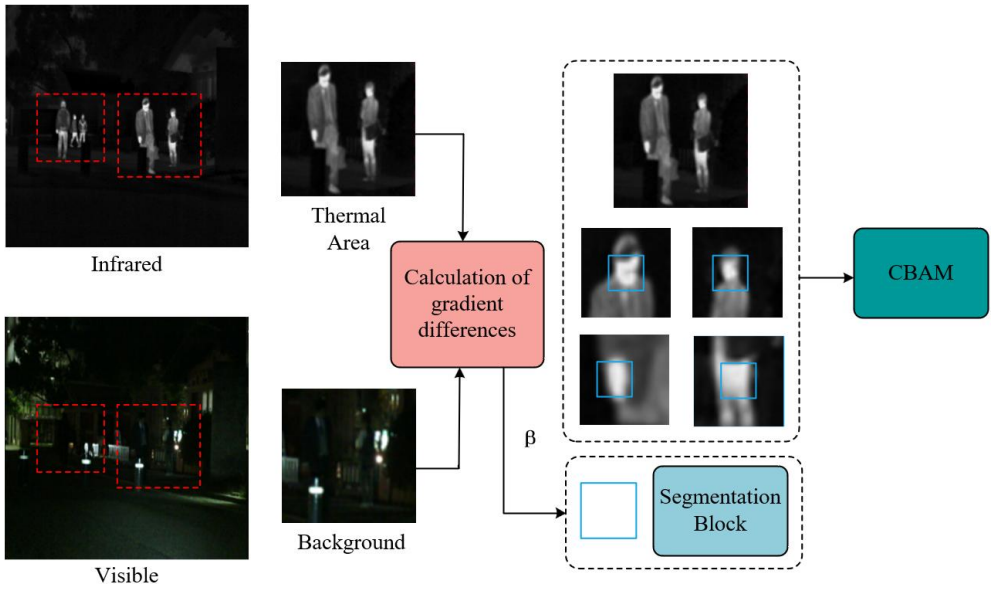


Fig.5. Negative Value Calibration Diffusion Attention Mechanism

3 Network Architecture Design

3.1 PTS Generator Module

To address overflow, boundary blurring, and brightness contamination in high-temperature regions during infrared-visible image fusion, we design a Precise Thermal Source (PTS) segmentation module, as illustrated in Fig. 4. The module takes the infrared image as input and first extracts multi-scale features through convolution and downsampling. The attention range is then determined by the Negative Value Calibration Diffusion mechanism (Fig. 5), which restricts attention to a $W \times W$ local window to reduce computational cost.Within this local region, a lightweight Convolutional Block Attention Module (CBAM) [14] is introduced for feature enhancement. By applying channel and spatial attention, the module selectively strengthens salient responses and highlights thermal-source regions. A Sigmoid activation is then used to generate a normalized thermal saliency mask.

During training, the PTS module is jointly optimized with the fusion generator. The segmentation branch is supervised by a thermal saliency loss composed of binary cross-entropy and a learnable β-related term, which helps improve thermal-region localization while preserving brightness in low-temperature areas. The resulting segmentation output, shown in Fig. 6, exhibits enhanced saliency in thermal targets and clear luminance separation from the background.

Meanwhile, the fusion branch is guided by multiple loss terms, including structural similarity (SSIM), gradient preservation (edge loss), and pixel reconstruction loss. This combined optimization strategy effectively suppresses highlight overflow and improves contrast, structural fidelity, and overall visual quality in the fused image.

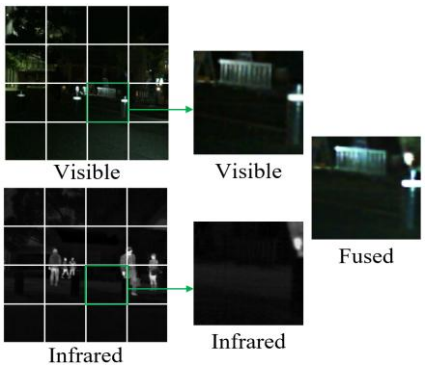


Fig.6. 00735N Segmentation Result

3.2 Multi-Channel Discriminator with Weighted Cross-Attention

To constrain the generated fused image to approximate the real distribution in both the intensity domain (RGB) and the high-frequency structural domain (gradient), this paper designs a dual-branch discriminator composed of a main RGB branch and a gradient-aware branch. Built on a patch-based discrimination framework, the discriminator further incorporates a weighted cross-attention network (CA-Net). This design jointly models global contrast and color consistency while emphasizing high-frequency information such as edges and textures, thereby improving the clarity and structural fidelity of the fused results.

The main discriminator branch takes a three-channel image $X \in R^{H \times W \times 3}$ as input. It employs hierarchical downsampling blocks constructed by stacking multiple 4×4 convolutional layers with stride 2, each followed by Instance Normalization.

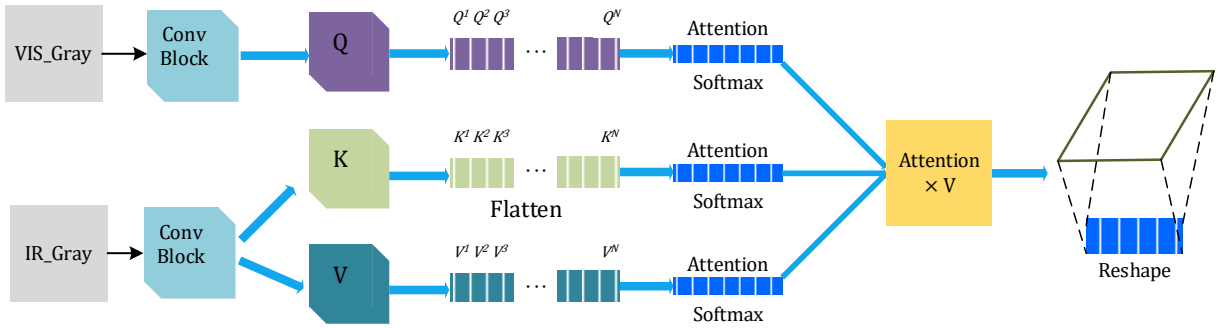


Fig.7. CA-Net Network Architecture

and a LeakyReLU activation. The process culminates in a 3×3 convolution that maps the features to a single-channel response map $D_{RGB}(x) \in R^{H' \times W'}$, achieving patch-wise authenticity discrimination.Building upon this foundation, a gradient-aware branch is introduced to enhance supervision over high-frequency details. The input is first converted to grayscale using BT.601 coefficients, as shown in the following equation:

$$X^{gray} = 0.299X_R + 0.587X_G + 0.114X_B \quad (15)$$

then extracted using Sobel operators:

$$G = \sqrt{G_x^2 + G_y^2 + \epsilon} \quad (16)$$

$$G_x = X^{gray} * S_x, \quad G_y = X^{gray} * S_y$$

where S_x and S_y denote the horizontal and vertical Sobel kernels respectively, and ε is a numerical stability term set to $0.01 \times [G_x + G_y]$. The gradient magnitude map G is processed through a lightweight downsampling stack architecturally consistent with the main branch but with reduced parameters, producing the gradient response map $D_{grad}(G) \in R^{H' \times W'}$. The discriminator loss is formulated as:

$$L_D = E_{x \in X}\left[\phi\left(1 - D_{rgb}(x)\right) + \lambda_g \phi\left(1 - D_{grad}(G(x))\right)\right] + E_{\hat{x} \in Y}\left[\phi\left(1 + D_{rgb}(\hat{x})\right) + \lambda_g \phi\left(1 + D_{grad}(G(\hat{x}))\right)\right] \quad (17)$$

Adversarial Loss for the Generator:

$$L_G^{adv} = E_{\hat{x} \in Y}\left[\phi\left(-D_{rgb}(\hat{x})\right) + \lambda_g \phi\left(-D_{grad}(G(\hat{x}))\right)\right] \quad (18)$$

where $\phi(\cdot)$ denotes the adversarial criterion (e.g., $\phi(z) = max(0, 1+z)$ for hinge loss or $\phi(z) = z^2$ for LSGAN), and λ_g regulates the importance of gradient-domain adversarial learning. During training, L_{Gadv} can be combined with weighted reconstruction terms—such as SSIM, perceptual loss, and edge preservation loss—to simultaneously enhance global realism and local sharpness.

2.2.1 CA-Net Attention Module

During discriminator fusion, a dynamic weighting strategy guided by cross-modal attention is introduced to enable more effective extraction of complementary information from infrared and visible images, as illustrated in Fig. 7. First, the infrared and visible images are converted to grayscale and projected into embedded feature representations through convolutional mapping. The visible grayscale features are used as the Query (Q), while the infrared grayscale features serve as the Key (K) and Value (V). Cross-modal attention is then computed at the scale of the main discriminator branch output, as defined in Equation (19):

$$Attn(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (19)$$

where d denotes the embedding dimension and is used for normalization to maintain numerical stability. The resulting cross-modal contextual features are further processed by a convolutional layer followed by a Sigmoid activation, producing a single-channel spatial weight map $\gamma \in [0,1]^{B \times 1 \times h \times w}$, whose spatial resolution matches that of the discriminator logits. Guided by this weight map, the outputs of the RGB branch and the gradient branch are fused as

$$D_{\text{out}} = \gamma \cdot D_{\text{grad}} + (1 - \gamma) \cdot D_{\text{rgb}} \quad (20)$$

This cross-modal attention mechanism[15] enables the discriminator to perform adaptive discrimination based on the modality dominance in different regions: it enhances the constraint from the gradient branch in infrared-dominant areas (e.g., thermal targets or edges), while preserving more texture and detail in visible-dominant regions, thereby obtaining more discriminative supervisory signals.

4 Experimental Results and Analysis

The proposed method was evaluated on a workstation equipped with an AMD EPYC 9754 CPU and an NVIDIA RTX 4090D GPU. The model was implemented in Python using the PyCharm integrated development environment. Both the proposed method and all comparison methods were trained and evaluated in full precision.

4.1 Datasets

Publicly available MSRS (Multi-Spectral Road Scenarios) and M3FD datasets were used for evaluation. Both datasets are designed for multi-spectral image fusion and multi-modal perception tasks, and each contains registered pairs of infrared (IR) and RGB visible (VIS) images. Focusing on road and traffic scenes, they cover diverse weather and illumination conditions, including daytime, nighttime, fog, and rain, thereby reflecting sensor imaging characteristics in complex environments. All images have a unified resolution of 640×480 pixels and are carefully aligned to ensure spatial correspondence between the infrared and visible modalities.

In this study, the number of low-visibility samples in the original MSRS dataset was increased to 4,252 image pairs through cropping and sample generation. The test set was constructed by uniform interval sampling with a ratio of 20:1, and an additional 20 images were reserved for convergence monitoring after each training epoch.

4.2 Training Configuration

The maximum number of training epochs was set to 100. To avoid unnecessary training and overfitting, an early stopping strategy was adopted: training was automatically

terminated if no performance improvement was observed for five consecutive epochs.

4.3 Fusion Performance Evaluation Metrics

Information-theoretic metrics are mainly used to quantify the overall information content contained in fused images. Representative indicators include entropy (EN), mutual information (MI), average gradient (AG), and standard deviation (SD). Specifically, EN reflects the richness and distribution uniformity of image intensity, MI measures the amount of shared information between the fused image and the source images, SF and AG characterize detail sharpness and texture richness, and SD evaluates grayscale contrast. In general, higher values of these metrics indicate that the fused image preserves more useful information information [16].

Structural similarity metrics are used to evaluate how well the fused image preserves the structural and textural characteristics of the source images. Commonly used indicators include the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), mean squared error (MSE), and correlation coefficient (CC). Among them, SSIM assesses similarity from the perspectives of luminance, contrast, and structure; PSNR and MSE measure reconstruction differences between the fused and source images; and CC reflects the consistency of grayscale variation trends [17].

Perceptual quality metrics place greater emphasis on human visual perception and task-relevant characteristics. For example, visual information fidelity (VIF) evaluates information preservation based on natural scene statistics, while the gradient-based fusion performance index $Q_{AB/F}$ measures how effectively edge and gradient information from the source images is transferred to the fused result.

Because no single metric can fully characterize fusion quality, this study adopts a multi-metric evaluation protocol including EN, AG, SD, MI, SSIM, VIF, and $Q_{AB/F}$. These metrics provide a comprehensive assessment of the proposed method from the perspectives of information content, structural consistency, and perceptual quality. Such a multidimensional evaluation framework not only verifies whether the fused image effectively preserves complementary information from infrared and visible images, but also demonstrates its advantages in visual quality and task adaptability.

4.4 Results Analysis and Comparison

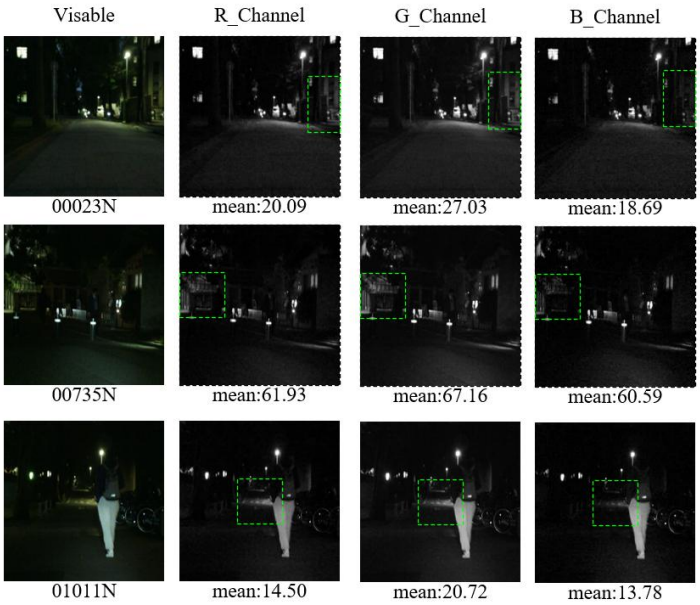


Fig.8. Grayscale Distribution of RGB Channels in MSRS Dataset

4.4.1 Color Balance Variation

In the infrared-guided analysis experiment, two representative samples, 00735N and 00023N, from the MSRS dataset were selected for visual comparison. These samples contain multiple thermal targets and weak thermal targets, respectively, allowing evaluation of the network's sensitivity to targets of different scales as well as its image reconstruction capability. After channel decomposition and statistical analysis of these two samples (Fig. 8), it was observed that the green channel had a significantly higher mean intensity than the red and blue channels in the original visible images. After further processing, the mean values of the three channels changed by -6.82% in R, -11.26% in G, and -5.79% in B relative to their original values, with the reduction in the green channel being substantially larger than those in the red and blue channels.

This result indicates that the proposed processing strategy effectively suppresses overall color cast. By reducing the dominance of the green channel, the distribution of the three channels becomes more balanced, thereby improving color reproduction and visual realism. Such adjustment not only weakens the excessive influence of the green channel on the overall tone of the image, but also enhances image stability and reliability for subsequent fusion and perception tasks.

Furthermore, taking sample 00735N as an example, the grayscale distribution during intermediate processing stages (Fig. 9) shows that the infrared responses are effectively constrained within the visible component. As a result, the final fused image achieves better consistency with both modalities in terms of visual quality and quantitative performance. This targeted suppression of high-mean channel intensity further improves color consistency, making the fused result more consistent with human visual expectations.

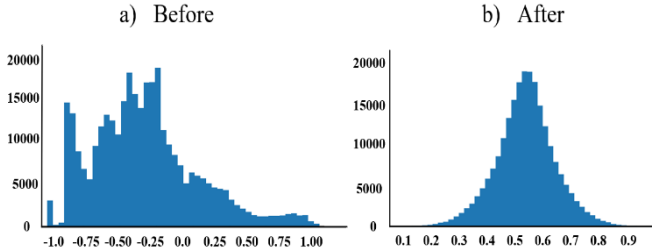


Fig.9. 00735N Post-Accurate-Segmentation Distribution

During inference, the input visible image is decomposed into three channels. The channel with the lowest grayscale mean is taken as the reference and normalized to 1.000. The deviation ratio of each remaining channel is then calculated as the ratio between its mean value and that of the reference channel. Under the adaptive hierarchical activation mechanism, the computational cost of image fusion varies with the chromatic deviation of the visible image, as summarized in Table 1.

For images in the MSRS dataset with larger inter-channel mean deviations, such as 00023N, 01011N, and 00735N, the deviation ratio of the green channel exceeded the adaptive threshold (initialized to 0.05 and dynamically updated through weighted averaging during processing). As a result, these channels were routed to the deep branch for further processing. In the final fused images, the dominance of the green channel was reduced, leading to a more visually balanced chromatic distribution.

By contrast, some images from the M3FD dataset, such as 04037 and 00359, exhibited much smaller deviations among the three visible channels, indicating a more uniform chromatic distribution. Consequently, their channel distributions remained largely consistent with the original images after fusion. For image 01443 from the same dataset, however, the larger deviations in the green and blue channels triggered deeper processing. Overall, the computational cost is positively correlated with the number of channels whose mean-deviation ratios exceed the adaptive threshold.

TABLE Ⅰ. The Impact of Mean Chromatic Deviation on Computational Load

Image	R_C/mean	G_C/mean	B_C/mean	R_C/bias	G_C/bias	B_C/bias	Flops/G
00023N	20.09	27.03	18.69	1.074	1.446	1.000	112
00735N	20.64	22.39	20.20	1.022	1.108	1.000	112
01011N	14.50	20.72	13.78	1.000	1.146	1.017	112
04037	86.99	88.56	88.35	1.000	1.018	1.017	157
01443	75.63	84.56	83.35	1.000	1.118	1.102	132

3.4.2 Comparative Evaluation

To validate the effectiveness of the proposed method, we conducted comparative experiments against typical fusion methods, including DDcGAN [18],[19] and SwinFusion [20], as well as several state-of-the-art methods with excellent performance, namely CrossFusion [21], DenseFusion [22], and FreeFusion [23]. To further verify the algorithm's reliability, representative images from the M3FD dataset were selected for testing.

A visual comparison on typical images from the MSRS dataset (Figs.10-11) reveals that, due to the dataset's generally simple background heat sources and low color saturation, the challenge of fusing detailed textures is relatively low. Consequently, the compared methods demonstrate similar performance in representing background features and pedestrian heat sources in the fusion results. Among them, Cross Fusion excels in scenarios with smaller thermal targets, as it enhances the contrast between the heat source and the background through cross-modal interaction, leading to superior performance for small thermal objects.Compared to these methods, the proposed fusion network not only achieves a better balance between infrared target enhancement and visible detail preservation but also shows consistent improvements across multiple objective metrics. This indicates that our method significantly enhances the information utilization and visual presentation of the fused image while ensuring structural fidelity. However, there remains room for improvement in the overall sharpness of the fused result. In multimodal image fusion and its downstream tasks (e.g., object detection), clear and sharp target contours can enhance the distinction between foreground and background, thereby improving the localization accuracy of regions of interest and the expressiveness of their features.

On the widely used MSRS dataset, the average metrics of our method, compared to the best-performing baseline, demonstrate significant gains: Entropy (EN) increased by 7.83%. These improvements in information-theoretic metrics indicate superior detail retention and richer grayscale level distribution in the fused images. Metrics for structural similarity also showed substantial improvements: Correlation Coefficient (CC) increased by 13.14%, and Structural Similarity (SSIM) increased by 4.23%, confirming that our method better preserves structural consistency and textural details from the source images. Meanwhile, in terms of perceptual quality metrics, Standard Deviation (SD) increased by 6.71%, and the gradient-based fusion performance $Q_{AB/F}$ increased by 14.96%, both reaching higher values, which further validates the advantages of our fusion results in visual quality and edge preservation. The average inference time was 0.0065s/image.

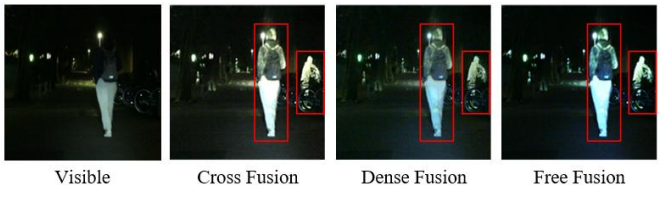


Fig.10 01011N Comparison of Results

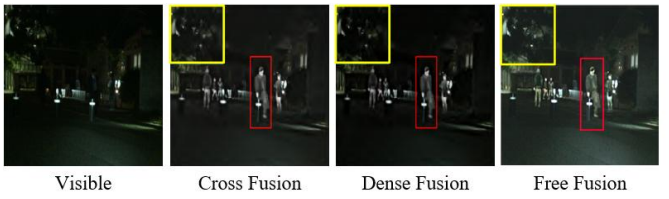


Fig.11 00735N Comparison of Results

TABLE II. Objective Evaluation Metrics for MSRS Fused Images

Image	EN	MI	$Q^{AB/F}$	AG	CC	SD	SSIM	VIF
01011N	7.913 6	3.957 2	0.709 2	4.842 4	0.649 8	48.634 3	0.811 7	0.799 7
00735N	7.950 1	3.994 5	0.742 3	4.544 7	0.673 6	48.323 5	0.800 1	0.791 1

TABLE III. Average Objective Evaluation Metrics on the MSRS Dataset

Method	EN	MI	$Q^{AB/F}$	AG	CC	SD	SSIM	VIF	Time/s
FusionGAN	6.350 0	2.159 6	0.312 3	2.587 5	0.547 6	27.435 3	0.674 7	0.442 0	0.015 1
DDcGAN	7.350 3	2.058 0	0.346 3	**5.722 5**	0.526 3	44.014 0	0.528 3	0.485 3	0.018 1
DenseFuse	6.369 0	2.436 3	0.348 3	3.967 3	0.595 9	26.829 0	0.733 7	0.572 7	0.010 4
SwinFusion	6.872 0	3.846 7	0.562 3	3.948 7	0.542 0	44.159 7	0.711 0	0.800 0	0.034 4
FreeFusion	7.228 0	4.134 2	0.674 3	4.032 6	0.528 6	45.382 0	0.772 9	0.769 0	**0.004 9**
CrossFuse	7.140 9	**4.462 7**	0.638 7	3.965 1	0.438 3	42.292 3	0.758 4	**0.918 5**	0.018 5
Ours	**7.920 8**	3.993 2	**0.734 3**	4.528 7	**0.674 2**	**48.428 4**	**0.804 0**	0.799 0	0.006 5

In the algorithm reliability test, the network trained on the MSRS dataset was applied to fuse image pairs from the M3FD dataset. The results in Fig.12 demonstrate that on the M3FD image pairs, while the CrossFusion method successfully enhances the infrared thermal targets within the region of interest (marked by the red box), its cross-modal interaction mechanism introduces excessively high contrast in the background areas outside the pedestrians. This leads to an overall image with overly strong sharpness. DDcGAN shows improvement in balancing the infrared target and visible details. However, constrained by its limited receptive field, it still suffers from local blurring in edge regions and deta , iled borders. DenseFusion achieves satisfactory highlighting of infrared targets with superior computational efficiency. Yet, the extensive use of skip connections causes the infrared components to disproportionately dominate the fusion result, resulting in significant loss of background details, particularly in high-frequency texture regions (marked by the yellow box), where its performance is suboptimal.

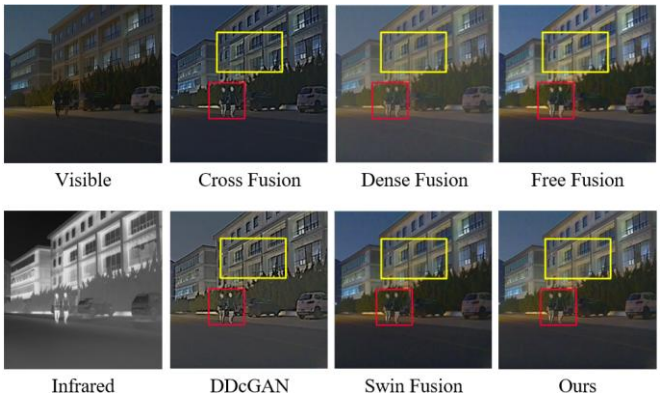


Fig.12. M3FD 04037 Comparison of Results

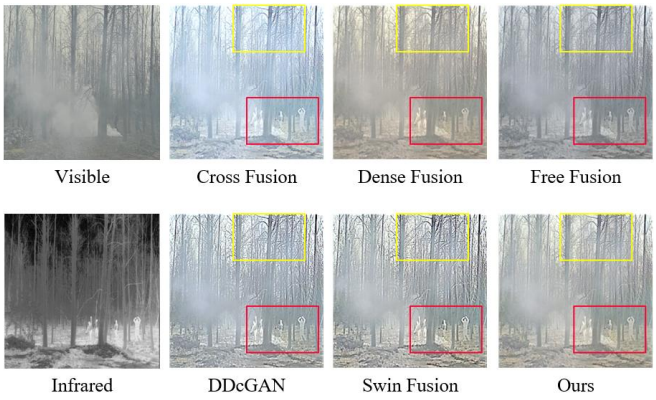


Fig.13. 01443 Comparison of Results

TABLE IV. Generalization Evaluation of the Proposed Method on the M3FD Dataset

Image	EN	MI	$Q^{AB/F}$	AG	CC	SD	SSIM	VIF
04037	7.964 5	3.993 2	0.763 7	4.534 6	0.696 7	49.543 4	0.891 6	0.846 3
01443	7.925 2	3.921 6	0.754 9	4.556 5	0.689 3	49.633 8	0.893 2	0.858 4

For the smoke-scene sample 01443 (Fig.13), the comparative results show that both DDcGAN and SwinFusion enhanced thermal targets but introduced over-sharpening in fine details. CrossFusion and DenseFusion exhibited noticeable color distortion and tended to amplify image noise. FreeFusion preserved details relatively well, but its fusion result was overly dominated by infrared information, leading to degraded visual quality metrics.

By contrast, under the challenging conditions of low-intensity thermal targets in smoke scenes, the proposed method effectively highlights salient infrared targets while preserving the structural details of the visible image. As a result, the fused image achieves better target distinguishability and higher background-fidelity preservation than the compared methods. Nevertheless, in terms of robustness, all methods showed some degree of color cast in this scenario, and the thermal enhancement effect was weaker than that observed on the MSRS dataset.

Quantitative results in Table 4 further demonstrate the effectiveness of the proposed method. Owing to the more balanced chromatic distribution in the M3FD dataset, the proposed method achieved stronger performance on structural and perceptual similarity metrics than on the MSRS dataset. Specifically, it reached an SSIM of 0.8916, a CC of 0.6834, and a VIF of 0.8373. These results indicate that the proposed method achieves a favorable balance between infrared saliency enhancement and visible texture preservation, providing clear advantages for both visual perception and downstream applications.

5. Conclusion

Experimental results on both the M3FD and MSRS datasets comprehensively demonstrate the effectiveness of the proposed method. From a qualitative perspective, the proposed approach achieves a favorable balance between infrared target enhancement and visible-detail preservation. In low-light scenes, the fused images clearly highlight thermal targets, such as pedestrians and vehicles, while avoiding unnatural over-enhancement or block-like artifacts. At the same time, rich visible details—including building contours, foliage structures, and ground textures—are well preserved, which significantly improves the naturalness and interpretability of the fused results.

From a quantitative perspective, compared with both traditional and state-of-the-art methods, the proposed fusion network consistently achieves superior performance across multiple evaluation metrics, including SSIM, CC, and SD. These results indicate that the proposed method effectively improves information preservation and visual quality while maintaining structural fidelity. Overall, this work provides strong support for practical applications of infrared-visible image fusion.

References

[1] Yu Liu, Xun Chen, Hu Peng, Zengfu Wang, “Multi-focus image fusion with a deep convolutional neural network,” Information Fusion, vol. 36, 2017, pp. 191–207, doi: 10.1016/J.INFFUS.2016.12.001.

[2] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, Junjun Jiang, “FusionGAN: A generative adversarial network for infrared and visible image fusion,” Information Fusion, vol. 48, 2018, pp.11–26, doi: 10.1016/J.INFFUS.2018.09.004

[3] John J. Lewis, Robert J. O’Callaghan, Stavri G. Nikolov, David R. Bull, Nishan Canagarajah, “Pixel- and region-based image fusion with complex wavelets,” Information Fusion, vol. 8, no. 2, 2005, pp. 119–130, doi: 10.1016/J.INFFUS.2005.09.006.

[4] Jiayi Ma, Chen Chen, Chang Li, Jun Huang “Infrared and visible image fusion via gradient transfer and total variation minimization,” Information Fusion, vol. 31, 2016, pp. 100–109, doi: 10.1016/J.INFFUS.2016.02.001.

[5] Li Hui, Wu Xiao-Jun, “DenseFuse: A Fusion Approach to Infrared and Visible Images,” IEEE Transactions on Image Processing, vol. 28, no. 5 (2019), pp. 2614–2623, doi: 10.1109/TIP.2018.2887342.

[6] Hui Li, Xiao-Jun Wu, Josef Kittler, “RFN-Nest: An end-to-end residual fusion network for infrared and visible images”, Information Fusion, vol.73, 2021, pp.72-86, doi:10.1016/J.INFFUS.2021.02.023.

[7] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, Jiangshe Zhang, “DIDFuse: Deep image decomposition for infrared and visible image fusion,” Information Fusion, vol. 64, 2020, pp. 27-39 ,doi: 10.48550/arXiv2003.09210.

[8] Xu Han, Ma Jiayi, Jiang Junjun, Guo Xiaojie, Ling Haibin, “U2Fusion: A Unified Unsupervised Image Fusion Network,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, 2020, pp. 502–518, doi: 10.1109/TPAMI.2020.3012548.

[9] Hu, Jie, Li Shen, Gang Sun, “Squeeze-and-Excitation Networks.”Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141, doi: 10.1109/TPAMI.2019.2913372.

[10] Jonathan Ho, Ajay N. Jain, Pieter Abbeel, “Denoising Diffusion Probabilistic Models,”in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp.6840–6851, doi: 10.48550/arXiv.2006.11239.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,eg.al, “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.

[12] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh, “Hierarchical question-image co-attention for visual question answering,” in Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 289–297, doi: 10.48550/arXiv.1606.00061.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, eg.al, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon, “CBAM: Convolutional block attention module,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[15] Zihuan Zhu, Xiaou Song, Wei Cui, Qi Fengli, “A Review on Target Detection in Visible Light-Infrared Image Fusion” [J]. Computer Engineering and Applications, 2025, 61 (17), pp. 17-32.

[16] Guihong Qu,,Dali Zhang, Pingfan Yan, “Information measure for performance of image fusion”[J].Electronics Letters,2002,38(7), pp.313-315, doi: 10.1049/el:20020213.

[17] Xydeas C ,PetroviU+0107 V .Objective image fusion performance measure[J].Electronics Letters, 2000, 36(4), pp.308-309, doi: 10.1049/el:20000267.

[18] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, Xiao-Ping Zhang, “DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion” in IEEE Transactions on Image Processing, 2020,vol. 29, pp. 4980-4995, doi: 10.1109/TIP.2020.2977576 .

[19] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, Yong Ma, “SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13018–13029, doi: 10.1109/CVPR52688.2022.01268.

[20] Li Hui, Xiaojun Wu, “CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach,” Information Fusion, vol. 103, art. no. 102147, 2024, doi: 10.1016/j.inffus.2024.102147.

[21] Pengwei Liang, Junjun Jiang, Xianming Liu, Jiayi Ma,“Fusion from Decomposition: A self-supervised decomposition approach for image fusion,” in Computer Vision – ECCV 2022, vol. 13680, Lecture Notes in Computer Science, pp. 719–735, doi: 10.1007/978-3-031-20071-7_42.

[22] Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, Huchuan Lu, "FreeFusion: Infrared and Visible Image Fusion via Cross Reconstruction Learning," IEEE transactions on pattern analysis and machine intelligence PP.(2025), doi: 10.1109/TPAMI.2025.3482689.

[23] Jiayi Ma,Yong Ma, Chang Li, “Infrared and visible image fusion methods and applications: A survey” Information Fusion 45.(2019),pp.153-178, doi: 10.1016/j.inffus.2018.02.004.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration: This work was supported by the Natural Science Foundation of Heilongjiang Province(PL2024F021).