Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.35085

LPGFusion: Enhancing Spatial Saliency in Image Fusion via Coordinate Positional Encoding and Cross-Modal Gating

Ruijie He

Guohao Honors College, Tongji University, Shanghai, China
2353904@tongji.edu.cn

Abstract. Infrared and visible image fusion aims to integrate thermal saliency from infrared images with texture and structural details from visible images. However, existing fusion methods often weaken spatial positional information during feature compression, resulting in blurred target boundaries and incomplete texture preservation. To address this issue, this paper proposes LPGFusion, a light-aware position-gated fusion network. The method employs dual-branch encoders to extract modality-specific features and introduces coordinate positional encoding to preserve horizontal and vertical spatial cues. A light perception network provides global illumination guidance, while a cross-modal gating module performs adaptive modality selection at each spatial location. In addition, Sobel edge loss and multi-scale structural similarity loss are incorporated to enhance edge fidelity and multi-scale structural consistency. Experiments on the MSRS dataset show that LPGFusion achieves superior or competitive performance against thirteen representative methods. Ablation studies verify the complementary contributions of the proposed modules, and qualitative YOLOv11 results further demonstrate the effectiveness of LPGFusion for downstream tasks.

Keywords: infrared and visible image fusion, coordinate positional encoding, cross-modal gating, multi-scale structural similarity, edge preservation

1. Introduction

Infrared and visible image fusion is a vital multi-modal task widely utilized in fields such as autonomous driving, intelligent surveillance, and robotic perception [1-4]. While visible images provide rich textures and environmental details consistent with human vision, their quality degrades significantly under adverse lighting or poor weather conditions. In contrast, infrared sensors capture thermal radiation to highlight salient targets like pedestrians and vehicles in low-light environments, though they often lack semantic background and suffer from blurred edges. By leveraging these complementary characteristics, image fusion integrates the strengths of both modalities to generate more informative representations with enhanced target saliency and background clarity.

In recent years, deep learning-based fusion methods have become dominant. DenseFuse adopts an encoder-decoder architecture to extract deep features and reconstruct fused images [5]. FusionGAN introduces adversarial learning to encourage the fused image to preserve infrared target intensity while maintaining visible textures [6]. GAN-FM uses full-scale skip connections and dual

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.35085

Markovian discriminators to strengthen local and global feature preservation [7]. CUFD decomposes image features into common and unique components for visible and infrared image fusion [8]. PIAFusion further introduces an illumination-aware mechanism to adaptively balance infrared and visible information under different lighting conditions [9]. Recent methods, such as MaeFuse, RCAFusion, and DCEvo, further explore pretrained feature transfer, cross-modal attention, and task-oriented evolutionary optimization for infrared-visible fusion [10-12].

Although these methods have improved the performance of infrared-visible image fusion, an important limitation still exists. Many fusion modules compress spatial information before or during fusion. For example, global pooling, channel-wise attention, and vectorized feature descriptors capture the global importance of channels, but weaken the exact spatial positions of targets and textures. However, spatial information is essential in infrared and visible image fusion. The infrared response of a pedestrian is important at the pedestrian region, while visible textures such as road lines, buildings, trees, and background contours are also spatially distributed. If the fusion process loses spatial sensitivity, the final image suffers from blurred target boundaries, weakened thermal saliency, or missing visible details.

To address these challenges, this paper proposes LPGFusion, a Light-aware Position-Gated network that enhances the PIAFusion framework [9] by prioritizing local, position-specific modality importance over global weighting. The architecture integrates coordinate positional encoding to capture direction-aware spatial features [13] and a cross-modal gating mechanism—guided by an illumination-aware light perception network—to adaptively assign spatial weights. Furthermore, the training process is optimized with Sobel edge and MS-SSIM losses, ensuring superior boundary preservation and multi-scale structural fidelity in the reconstructed fused images [14].

The main contributions of this paper are summarized as follows:

1. A light-aware position-gated fusion framework is proposed for infrared and visible image fusion. The framework explicitly enhances spatial information before feature fusion and reduces the information loss caused by spatial compression.

2. A cross-modal gating module is designed to adaptively select infrared or visible features at each spatial location. This enables the network to preserve infrared salient targets and visible background textures simultaneously.

3. A Sobel edge loss and an MS-SSIM loss are introduced into the training objective. These two losses improve edge preservation and multi-scale structural fidelity.

4. Extensive comparison and ablation experiments are conducted. Experimental results show that LPGFusion achieves superior performance on most quantitative metrics. Qualitative YOLOv11 detection results further demonstrate its usefulness for downstream perception tasks.

2. Materials and methods

2.1. Overall framework

Given a registered infrared image I_{ir} and a visible image I_{vis}, the objective of infrared-visible image fusion is to generate a fused image I_f that contains both infrared thermal target information and visible texture details. Following common fusion practice and PIAFusion [9], the visible image is transformed from RGB space to YCbCr space. The luminance channel I_y is fused with the infrared image I_{ir}, while the chrominance channels of the visible image are retained for color reconstruction. The proposed LPGFusion mainly consists of four parts: dual-branch feature extraction, coordinate positional encoding, light-aware cross-modal gated fusion, and image

DOI: 10.54254/2755-2721/2026.35085

reconstruction. The visible luminance image and infrared image are first fed into two independent encoders:

$$F_{\text{vis}} = E_{\text{vis}}\left(I_{\text{vis}}^{Y}\right), \quad F_{\text{ir}} = E_{\text{ir}}\left(I_{\text{ir}}\right) \tag{1}$$

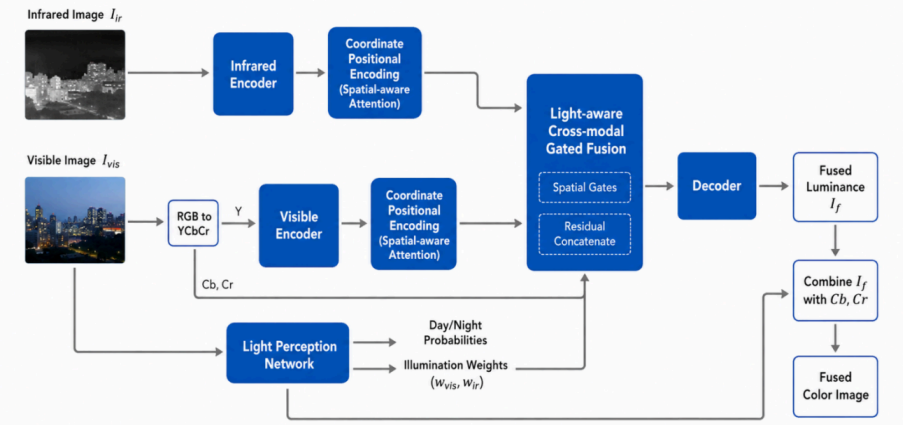

Figure 1. Overall framework of the proposed LPGFusion

where E_{vis} and E_{ir} are the visible and infrared encoder branches. In the implementation, each branch contains a 3x3 convolutional ConvBlock stem and a second ConvBlock encoder. With base_channels = 32, the stem output has 32 channels and the encoder output has 64 channels.

Coordinate positional encoding is then applied to the two modality features:

$$\widetilde{F}_{\text{vis}} = CA\left(F_{\text{vis}}\right), \quad \widetilde{F}_{\text{ir}} = CA\left(F_{\text{ir}}\right) \tag{2}$$

where CA(.) denotes the coordinate positional encoding module.

For local modality selection, the enhanced features are sent to the light-aware cross-modal gate. The gated path and the residual concatenation path are combined as:

$$F_{\text{fused}} = G_{\text{vis}} \otimes \widetilde{F}_{\text{vis}} + G_{\text{ir}} \otimes \widetilde{F}_{\text{ir}} + \psi\left(\left[\widetilde{F}_{\text{vis}}, \widetilde{F}_{\text{ir}}\right]\right) \tag{3}$$

where, G_{vis} and G_{ir} are the visible and infrared spatial gate maps generated by a softmax operation, $\psi(.)$ is the convolutional transformation of the concatenated features, [.,.] denotes channel-wise concatenation, and $\otimes$ denotes element-wise multiplication. Finally, the decoder reconstructs the fused feature into the fused image:

$$\widehat{I}_f = D\left(F_{\text{fused}}\right) \tag{4}$$

The complete architecture is shown in Fig. 1. Compared with fusion methods that only rely on global feature weights, LPGFusion performs position-sensitive modality selection. This allows the

DOI: 10.54254/2755-2721/2026.35085

model to preserve infrared targets in salient regions and visible textures in background regions.

The infrared image and visible luminance image are input into two encoder branches. Coordinate positional encoding is used before fusion. A light perception network and a cross-modal gate jointly guide feature fusion. The decoder reconstructs the final fused image.

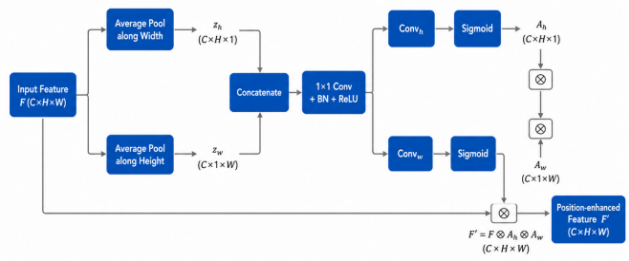

Figure 2. Coordinate positional encoding module

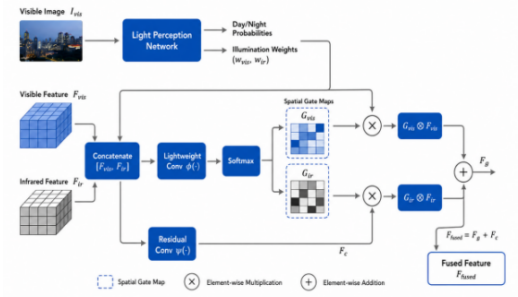

Figure 3. Light-aware cross-modal gated fusion module

2.2. Coordinate positional encoding module

Spatial information is crucial for infrared and visible image fusion. Infrared images mainly provide thermal targets, which are strongly related to specific object positions. Visible images provide texture and contour information, which is also spatially distributed. However, conventional channel attention often compresses a two-dimensional feature map into a one-dimensional vector through global average pooling. Although such an operation estimates channel importance, it weakens accurate spatial location information.

To preserve spatial information during feature fusion, LPGFusion introduces coordinate positional encoding [13], as shown in Fig. 2. Given an input feature map X in $\mathbb{R}^{C\times H\times W}$, two directional descriptors are generated by average pooling along the horizontal and vertical directions:

$$z_h(c,h) = \frac{1}{W}\sum_{w=1}^{W} X(c,h,w), \quad z_w(c,w) = \frac{1}{H}\sum_{h=1}^{H} X(c,h,w) \tag{5}$$

The two descriptors preserve vertical and horizontal position information, respectively. They are concatenated and transformed by a shared 1x1 convolution and an HSwish activation:

$$u = \delta\left(\mathrm{Conv}_{1\times 1}\left([z_h, z_w]\right)\right) \tag{6}$$

Then u is split into u_h and u_w, and two direction-aware attention maps are generated:

$$A_h = \sigma\left(\mathrm{Conv}_h(u_h)\right), \quad A_w = \sigma\left(\mathrm{Conv}_w(u_w)\right) \tag{7}$$

$$\widetilde{X} = X \otimes A_h \otimes A_w \tag{8}$$

DOI: 10.54254/2755-2721/2026.35085

where Conv_h and Conv_w are learnable 1x1 convolutions, δ denotes HSwish activation, σ denotes sigmoid activation, and $\widetilde{X}$ is the position-enhanced feature.

This mechanism preserves long-range dependencies along one spatial direction while maintaining precise positional information along the other direction. Therefore, it is suitable for infrared and visible image fusion, where object positions, edges, and textures are important. By applying this module before feature fusion, LPGFusion reduces the spatial information loss caused by feature compression.

2.3. Light-aware cross-modal gated fusion module

Different illumination conditions require different fusion preferences. In daytime scenes, visible images usually contain sufficient texture and semantic background information. In nighttime or low-light scenes, infrared images are more reliable for highlighting pedestrians and vehicles. Therefore, a fixed fusion rule cannot adapt well to complex scenes.

To address this problem, a light perception network estimates the illumination condition from the visible RGB image, as shown in Fig. 3. The network output is converted into probabilities by softmax:

$$(p_{\mathrm{day}}, p_{\mathrm{night}}) = softmax\left(C\left(I_{\mathrm{vis}}\right)\right) \tag{9}$$

The visible and infrared modality weights are computed as:

$$\omega_{\mathrm{vis}} = \frac{\max(p_{\mathrm{day}}, \tau_{\mathrm{vis}})}{Z}, \quad \omega_{\mathrm{ir}} = \frac{\max(p_{\mathrm{night}}, \tau_{\mathrm{ir}})}{Z}, \quad Z = max\left(p_{\mathrm{day}}, \tau_{\mathrm{vis}}\right) + max\left(p_{\mathrm{night}}, \tau_{\mathrm{ir}}\right) \tag{10}$$

where, τ_{vis} and τ_{ir} are lower-bound thresholds, set to 0.45 and 0.15 in the code, respectively. They prevent either modality from being completely suppressed.

Global illumination weights alone are not sufficient because different regions in the same image require different modality preferences. Therefore, LPGFusion introduces a cross-modal gating mechanism. The enhanced visible and infrared features are concatenated, and a lightweight convolutional network predicts two spatial gate maps:

$$[G_{\mathrm{vis}}, G_{\mathrm{ir}}] = softmax\left(\phi\left(\left[\widetilde{F}_{\mathrm{vis}}, \widetilde{F}_{\mathrm{ir}}\right]\right)\right) \tag{11}$$

The final fused feature is calculated using the compact formulation in (3). In the implementation, φ(.) consists of two 1x1 convolutional layers with LeakyReLU activations, while ψ(.) comprises a 3x3 convolution followed by LeakyReLU. The softmax operation normalizes the two gating weights at each spatial location so that their sum equals one.

Through this design, LPGFusion combines global illumination guidance and local spatial modality selection. The model adaptively emphasizes infrared features in target regions and visible features in texture-rich background regions.

DOI: 10.54254/2755-2721/2026.35085

2.4. Loss function

The loss function is designed to preserve illumination-adaptive content, salient regions, global content, edges, multi-scale structure, intensity, and brightness. According to the training code, the overall objective includes seven weighted terms: illumination loss, auxiliary reconstruction loss, content loss, edge loss, MS-SSIM loss, intensity loss, and brightness lower-bound loss.

2.4.1. Illumination-aware reconstruction loss

The illumination-aware target is generated from the classifier-guided modality weights:

$$T_{\text{illum}} = \omega_{\text{vis}} I_{\text{vis}}^{Y} + \omega_{\text{ir}} I_{\text{ir}}, \quad L_{\text{illum}} = \rho\left(\widehat{I}_f, T_{\text{illum}}\right) \tag{12}$$

where the Charbonnier penalty is defined as, with ε set to $1\text{x}10^{-6}$:

$$\rho(x, y) = mean\sqrt{(x-y)^2 + \varepsilon} \tag{13}$$

The code also uses a soft infrared-saliency mask:

$$M_{\text{ir}} = \sigma\left(\frac{I_{\text{ir}} - I_{\text{vis}}^{Y} - \delta_{\text{aux}}}{\tau_{\text{aux}}}\right), \quad M_{\text{vis}} = 1 - M_{\text{ir}} \tag{14}$$

The auxiliary and content losses are:

$$L_{\text{aux}} = \rho_{M_{\text{ir}}}\left(\widehat{I}_f, I_{\text{ir}}\right) + \rho_{M_{\text{vis}}}\left(\widehat{I}_f, I_{\text{vis}}^{Y}\right), \quad L_{\text{cont}} = \rho\left(\widehat{I}_f, M_{\text{ir}} I_{\text{ir}} + M_{\text{vis}} I_{\text{vis}}^{Y}\right) \tag{15}$$

where δ_{aux} =0.10 and τ_{aux} =0.03.

2.4.2. Sobel edge loss

Edges are important for both visual quality and downstream object detection. Infrared images provide target contours, while visible images provide texture boundaries. To enhance edge preservation, LPGFusion introduces a Sobel edge loss.

A target edge map is constructed by selecting the stronger weighted response:

$$T_{\text{edge}} = \frac{\max\left(\alpha_{\text{vis}} S\left(I_{\text{vis}}^{Y}\right), \alpha_{\text{ir}} S\left(I_{\text{ir}}\right)\right)}{\max\left(\alpha_{\text{vis}}, \alpha_{\text{ir}}\right)}, L_{\text{edge}} = \rho\left(S\left(\widehat{I}_f\right), T_{\text{edge}}\right) \tag{16}$$

where α_{vis} =1.2 and α_{ir} =0.8 are the edge weights used in the code,S(.) denote the Sobel gradient magnitude operator.

DOI: 10.54254/2755-2721/2026.35085

This loss encourages the fused image to preserve both infrared object boundaries and visible texture edges. Compared with ordinary gradient loss, Sobel loss provides stronger directional edge constraints.

2.4.3. MS-SSIM loss

Single-scale SSIM measures structural similarity under one scale. Fusion images contain both small local textures and large salient targets. Therefore, a multi-scale structural constraint is more suitable.

The structural target used by the code is the pixel-wise maximum between the visible luminance image and the gain-adjusted infrared image:

$$T_{\mathrm{ms}} = max\left(I_{\mathrm{vis}}^{Y}, \eta I_{\mathrm{ir}}\right), \quad L_{\mathrm{ms}} = \beta\left(1 - \mathrm{MS-SSIM}\left(\widehat{I}_f, T_{\mathrm{ms}}\right)\right) + (1-\beta)\rho\left(\widehat{I}_f, T_{\mathrm{ms}}\right) \tag{17}$$

where η=0.85 is the infrared intensity gain and β=0.84 controls the balance between MS-SSIM and pixel reconstruction.

This loss improves the structural fidelity of the fused image at different scales.

2.4.4. Intensity and brightness constraints

To prevent the fused image from becoming too dark, an intensity target is constructed with the same gain-adjusted maximum rule:

$$T_{\mathrm{int}} = \max\left(I_{\mathrm{vis}}^{Y}, \eta I_{\mathrm{ir}}\right),$$

$$L_{\mathrm{int}} = \rho\left(\widehat{I}_f, T_{\mathrm{int}}\right) \tag{18}$$

A brightness lower-bound loss is also used:

$$T_b = max\left(\mathrm{mean}\left(I_{\mathrm{vis}}^{Y}\right), \mu\,\mathrm{mean}\left(I_{\mathrm{ir}}\right)\right), \quad L_{\mathrm{bri}} = mean\left(\mathrm{ReLU}\left(T_b - \mathrm{mean}\left(\widehat{I}_f\right)\right)\right) \tag{19}$$

2.4.5. Total loss

The total loss is formulated as:

$$L = \lambda_{\mathrm{illum}} L_{\mathrm{illum}} + \lambda_{\mathrm{aux}} L_{\mathrm{aux}} + \lambda_{\mathrm{cont}} L_{\mathrm{cont}} + \lambda_{\mathrm{edge}} s_e L_{\mathrm{edge}} \tag{20}$$

$$+\lambda_{ms} L_{ms} + \lambda_{int} L_{int} + \lambda_{bri} L_{bri}$$

In (17), s_e is the edge-loss warm-up factor. The code uses the following seven loss weights:

DOI: 10.54254/2755-2721/2026.35085

$$(\lambda_{\text{illum}}, \lambda_{\text{aux}}, \lambda_{\text{cont}}, \lambda_{\text{edge}}, \lambda_{\text{ms}}, \lambda_{\text{int}}, \lambda_{\text{bri}}) = (1, 0.2, 1, 2, 1, 5, 2.5).$$

The first, second, and fourth weights are parsed from loss_weight=[1,0.2,2], while the other four are content weight, ms-ssim weight, intensity weight, and brightness weight. The Sobel edge loss and MS-SSIM loss are the two main loss-function improvements in this paper.

3. Experiments

3.1. Experimental settings

The proposed method is implemented based on the PIAFusion framework [9]. The training dataset is MSRS [15].The model is trained using the Adam optimizer. The initial learning rate is $5\text{x}10^{-4}$, and the learning rate is linearly decayed in the second half of training. The batch size is set to 2, and the number of training epochs is set to 30. The training crop size is set to 64x64. Horizontal flipping is used for data augmentation. Weight decay is set to 1x 10^{-5} , gradient clipping is set to 5.0, and dropout is set to 0.03 to stabilize the training process.

The proposed method is compared with thirteen representative infrared and visible image fusion methods, including CUFD [8], CSF [16], DenseFuse [5], DIVFusion [17], FusionGAN [6], GAN-FM [7], GANMcC [18], PIAFusion [9], PMGI [19], SDNet [20], SeAFusion [15], STDFusionNet [21], and U2Fusion [22].

Five objective evaluation metrics are used: entropy (EN), mutual information (MI), standard deviation (SD), visual information fidelity (VIF), and Q_{abf} [23]. EN measures the amount of information contained in the fused image. MI evaluates how much information is transferred from source images to the fused image. SD reflects image contrast and gray-level distribution. VIF measures visual information fidelity. Q_{abf} evaluates edge information preservation. For these metrics, larger values generally indicate better fusion performance.

3.2. Quantitative comparison

Table 1. Comparison between the proposed method and other fusion methods

Method	EN	MI	SD	VIF	$\mathbf{Q_{abf}}$
CUFD	6.056	2.983	33.986	0.584	0.435
CSF	5.785	3.454	29.155	0.658	0.472
DenseFuse	6.217	2.636	29.023	0.758	0.488
DIVFusion	5.954	3.023	31.512	0.688	0.461
FusionGAN	6.168	3.412	32.516	0.643	0.422
GAN-FM	6.203	3.211	28.735	0.598	0.411
GANMcC	6.125	2.507	26.341	0.625	0.298
PIAFusion	6.619	4.012	40.375	0.965	0.644
PMGI	5.964	2.788	32.556	0.601	0.407
SDNet	5.255	1.647	17.323	0.489	0.371
SeAFusion	6.651	4.031	41.843	0.969	0.675

DOI: 10.54254/2755-2721/2026.35085

Table 1. (continued)

STDFusionNet	5.649	3.444	33.685	0.591	0.496
U2Fusion	5.561	2.240	27.709	0.545	0.421
Ours	6.698	5.371	42.130	0.938	0.677

Table I shows that the proposed LPGFusion achieves the best performance on EN, MI, SD, and Q_{abf} . The EN value reaches 6.698, indicating that the proposed method preserves more image information. The MI value reaches 5.371, which is higher than PIAFusion and SeAFusion and demonstrates that LPGFusion transfers more complementary information from the source images to the fused image. The SD value of the proposed method is 42.130, which is also the highest among all compared methods, indicating stronger contrast and richer gray-level variations. The Q_{abf} value reaches 0.677, slightly higher than SeAFusion, showing that the proposed Sobel edge loss and position-gated fusion mechanism are effective for edge preservation. The VIF value of LPGFusion is 0.938, which is slightly lower than PIAFusion and SeAFusion. Overall, LPGFusion achieves the best or competitive performance on most metrics, proving the effectiveness of the proposed method.

3.3. Ablation study

To verify the individual and cumulative effectiveness of each proposed component, we conduct an ablation study based on the baseline PIAFusion framework [9]. Four configurations are designed: Baseline + Sobel Loss, Baseline + MS-SSIM Loss, Baseline + Position-Gated Attention, and the full LPGFusion model with all three improvements.

Table 2. Ablation results of the proposed components

Sobel Loss	MS-SSIM	PG-Atten.	EN	MI	SD	VIF	$\mathbf{Q_{abf}}$
√	×	×	6.568	5.035	42.023	0.922	0.661
×	√	×	6.639	4.531	42.123	0.577	0.633
×	×	√	6.648	5.647	41.615	0.884	0.674
√	√	√	6.698	5.371	42.130	0.938	0.677

Table II shows the distinct contributions of each module. The Sobel-loss configuration obtains a Q_{abf} value of 0.661 and an SD value of 42.023, demonstrating that dedicated edge supervision guides the network to preserve sharp target contours and fine visible textures. The MS-SSIM configuration increases EN and SD to 6.639 and 42.123, respectively, indicating that multi-scale structural constraints enrich image information and local contrast. The position-gated attention configuration achieves the highest MI score of 5.647 and a high Q_{abf} value of 0.674, proving that maintaining direction-aware spatial dependencies boosts information transfer from both source modalities. The full model achieves the top performance in EN (6.698), SD (42.130), VIF (0.938), and Q_{abf} (0.677). These results validate that the three proposed enhancements are mutually complementary and collectively essential for high-quality image fusion.

3.4. Downstream detection evaluation

To further analyze the practical value of the fused images, YOLOv11 is used to conduct object detection on fusion results [24]. The detected categories include person and car.

DOI: 10.54254/2755-2721/2026.35085

The qualitative detection results show that visible images provide rich background information in daytime scenes, but they miss pedestrians under low illumination. Infrared images highlight pedestrians and other thermal targets, but vehicle details and background structures are insufficient. Existing fusion methods improve detection performance to some extent, but missed detections or low-confidence boxes still appear in distant or low-contrast regions. The fused images generated by LPGFusion provide clearer object boundaries and more complete background structures. The Sobel edge loss strengthens target contours, while the MS-SSIM loss improves structural consistency. The position-gated attention mechanism further helps preserve the spatial locations of salient targets. As a result, YOLOv11 produces more stable bounding boxes and higher-confidence detection results in challenging scenes. It should be noted that the detection results are used only as qualitative evidence. The validation subset of the proposed method contains fewer images than the full comparison set of public fusion methods. Therefore, the detection results are not used as a strict quantitative comparison table. Instead, they demonstrate that LPGFusion benefits downstream perception tasks.



Figure 4. YOLOv11 detection results on different fusion images

4. Conclusion

This paper proposes LPGFusion, a light-aware position-gated network that enhances the PIAFusion framework by integrating coordinate positional encoding, cross-modal gated fusion, Sobel edge loss,

DOI: 10.54254/2755-2721/2026.35085

and MS-SSIM loss. The results demonstrate that preserving spatial information is essential for effective infrared and visible image fusion. Specifically, the coordinate positional encoding mechanism mitigates information loss caused by early feature compression by retaining horizontal and vertical positional cues, while the cross-modal gate offers stronger flexibility than global weighting through adaptive local modality selection. Furthermore, the loss-function improvements significantly bolster performance: Sobel edge loss directly supervises the edge response to preserve target boundaries and texture edges, while MS-SSIM provides multi-scale structural constraints that ensure fidelity in both local and global structures. Experimental benchmarks indicate that LPGFusion achieves state-of-the-art performance in EN, MI, SD, and Q_{abf}, with qualitative YOLOv11 detection results further validating its ability to provide clearer object boundaries and more stable responses for downstream perception. Although limitations remain regarding registration sensitivity and the reliance on global illumination guidance, LPGFusion proves effective for multi-modal image fusion and exhibits practical value for intelligent applications such as autonomous driving and surveillance.

References

[1] W. Ma, K. Wang, J. Li, Y. Qiao, Q. Li, and Y. Zhang, "Infrared and visible image fusion technology and application: A review, " Sensors, vol. 23, no. 2, Art. no. 599, 2023.
[2] A. Li, Z. Wang, F. Wang, Z. Liu, G. Yin, R. Fang, and K. Geng, "A novel semantic information perception architecture for extreme targets detection in complex traffic scenarios, " IEEE Transactions on Intelligent Vehicles, vol. 10, no. 5, pp. 3238-3250, 2025, doi: 10.1109/TIV.2024.3450201.
[3] C. Li, S. Zhou, Z. Liu, W. Zhang, and T. Wu, "TransAUAV: A transformer-enhanced RGB-infrared fusion network for anti-UAV detection, " IEEE Transactions on Aerospace and Electronic Systems, vol. 62, pp. 3498-3508, 2026, doi: 10.1109/TAES.2025.3646996.
[4] N. Shankar, P. Ladosz, and H. Yin, "CLEAR-IR: Clarity-enhanced active reconstruction of infrared imagery, " arXiv: 2510.04883, 2025.
[5] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images, " IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2614-2623, 2019, doi: 10.1109/TIP.2018.2887342.
[6] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion, " Information Fusion, vol. 48, pp. 11-26, 2019.
[7] H. Zhang, J. Yuan, X. Tian, and J. Ma, "GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators, " IEEE Transactions on Computational Imaging, vol. 7, pp. 1134-1147, 2021.
[8] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "CUFD: An encoder-decoder network for visible and infrared image fusion based on common and unique feature decomposition, " Computer Vision and Image Understanding, vol. 218, Art. no. 103407, 2022.
[9] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware, " Information Fusion, vol. 83-84, pp. 79-92, 2022.
[10] J. Li, J. Jiang, P. Liang, J. Ma, and L. Nie, "MaeFuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training, " IEEE Transactions on Image Processing, vol. 34, pp. 1340-1353, 2025, doi: 10.1109/TIP.2025.3541562.
[11] A. Li, G. Yin, Z. Wang, and J. Liang, "RCAFusion: Cross Rubik cube attention network for multi-modal image fusion of intelligent vehicles, " in Proc. IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Republic of Korea, 2024, pp. 2848-2854, doi: 10.1109/IV55156.2024.10588756.
[12] J. Liu, B. Zhang, Q. Mei, X. Li, Y. Zou, Z. Jiang, L. Ma, R. Liu, and X. Fan, "DCEvo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion, " in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 2226-2235.
[13] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design, " in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13713-13722.
[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment, " in Proc. 37th Asilomar Conference on Signals, Systems and Computers, 2003, pp. 1398-1402.

DOI: 10.54254/2755-2721/2026.35085

[15] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, " Information Fusion, vol. 82, pp. 28-42, 2022.
[16] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion, " IEEE Transactions on Computational Imaging, vol. 7, pp. 824-836, 2021.
[17] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion, " Information Fusion, vol. 91, pp. 477-493, 2023.
[18] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion, " IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-14, 2021, doi: 10.1109/TIM.2020.3038013.
[19] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, " in Proc. AAAI Conference on Artificial Intelligence, 2020, pp. 12797-12804.
[20] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, " International Journal of Computer Vision, vol. 129, pp. 2761-2785, 2021.
[21] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection, " IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-13, 2021.
[22] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network, " IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 502-518, 2022.
[23] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure, " Electronics Letters, vol. 36, no. 4, pp. 308-309, 2000.
[24] Ultralytics, "YOLO11: Real-time object detection model, " 2024. [Online]. Available: Ultralytics documentation.