International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Multimodal Image Fusion Network Based on Class Activation and Multiscale Edge Gradients

Shenglin Yang

North China Electric Power University, School of Control and Computer Engineering, Beijing, 102206, China
Email: *3877258648[at]qq.com*

Abstract: ***Image fusion has garnered significant attention for effectively addressing the inadequate information representation of single-modality images. However, existing methods often overemphasize visual effects while neglecting the requirements of downstream high-level vision tasks. To tackle this issue, this paper proposes a Semantic Information Driven Multimodal Image Fusion network (SIDM-Fusion). First, a semantic-driven fusion framework is constructed to dynamically guide the fusion process by establishing the correspondence between modal features and class activation weights. Second, a Multiscale Edge Gradient Block (MEGB) is designed to adaptively extract multiscale local features while reinforcing edge information in combination with the Sobel operator. Finally, a Semantic Prior Classification Network (SPC-Net) is proposed to extract multi-level semantic information by establishing long-range dependencies. By introducing class activation weights into the feature fusion process, an adaptive and fully learnable fusion rule is achieved, avoiding the need for manual design. Experimental results demonstrate that, compared with current state-of-the-art methods, the proposed network achieves superior fusion performance across multiple datasets, with improvements of 22% in Average Gradient (AG) and 14% in Mutual Information (MI).***

Keywords: Multiscale features, Infrared and visible images, Multimodal image fusion, Deep learning**.**

1. Introduction

Image fusion integrates multiple images or image sequences of different modalities obtained from various sensors to generate a fused image with a more comprehensive scene representation[1]. This effectively addresses the issue that a single-modality image usually fails to provide comprehensive scene information. As a crucial branch of multi-sensor image fusion, infrared (IR) and visible (VIS) image fusion can complement the characteristics of IR images (which feature prominent salient targets and robustness to illumination changes, but suffer from heavy noise and low contrast) and VIS images (which possess high resolution, rich texture details, and structural information, but are vulnerable to harsh conditions). Consequently, this complementary process yields informative fused images, compensating for the limitations of insufficient information representation in a single modality. It has been widely applied in various fields such as intelligent traffic monitoring, UAV disaster monitoring, and security surveillance, providing strong support for improving system performance and reliability[2, 3].

Traditional image fusion typically employs various mathematical transformations to design fusion rules. Although these methods possess strong interpretability, their feature extraction relies heavily on manual design, which limits the capability to capture image information and fails to account for the intrinsic differences between images. Meanwhile, as modern vision tasks become increasingly complex, the corresponding mathematical transformations have grown complicated, resulting in high computational costs. Therefore, the traditional approach of manually designing fusion rules can no longer meet current practical application demands. With the rapid development of deep learning, artificial neural networks map images into high-dimensional spaces to automatically extract features. Compared to traditional methods, these extracted features are more comprehensive, profound, and highly exploitable, demonstrating distinct advantages and laying the foundation for the rapid advancement of image fusion methods.

Currently, deep learning-based image fusion methods mainly include Generative Adversarial Network (GAN)-based methods[4, 5] and autoencoder-based methods[6, 7]. GAN-based methods effectively balance the feature distributions of different modality images through adversarial training; however, their training processes are often unstable, and the features of the fused images tend to bias toward a specific modality. Li et al.[8] proposed a novel fusion framework, DenseFuse (Dense Feature Fusion for Image Fusion), which employs a pre-trained autoencoder for image fusion. Li et al.[9] also proposed an end-to-end residual fusion network (RFN-Nest) that adopts a Nest connection structure and uses neural networks to fuse feature information. Zhang et al.[10] proposed a feature compression and multi-dimensional decomposition network (SDNet) to improve fusion performance through adaptive decision-making. Tang et al.[11] proposed an illumination-aware progressive image fusion network capable of determining loss function weights. Xu et al.[12] proposed a unified unsupervised image fusion (U2Fusion) network, consolidating different fusion tasks into a single framework. While these methods have achieved excellent results, they blindly pursue the visual effects of the fused images during the fusion process, thereby neglecting the practical requirements of downstream high-level vision tasks.

To address this issue, Tang et al. proposed a semantic-aware IR and VIS fusion (SeAFusion) network by attaching a segmentation model to integrate as much semantic information as possible, thereby enhancing the performance of high-level vision tasks on the fused images. Subsequently, Liu et al.[13] and Sun et al.[14] applied fusion networks from an object detection perspective to retain more semantic information. Segmentation models represented by SeAFusion have provided new perspectives for the practical application of image fusion and achieved significant breakthroughs in

meeting high-level vision task requirements. However, during the modeling process, they only consider the needs of high-level vision tasks and merely make improvements at the loss function level, ignoring the deep intrinsic connection between the fusion problem and the high-level vision tasks[15]. Furthermore, they completely rely on maximum selection strategies, which leads to an inability to accurately segment objects in extreme scenarios. Aiming at this problem, this paper proposes a Semantic Information Driven Multimodal Image Fusion (SIDM-Fusion) network for IR and VIS images. The main contributions are as follows:

1) **Designed a semantic-driven multimodal image fusion network:** Based on the class activation mapping mechanism, it establishes the correspondence between the features of each modality and the activation weights, utilizes semantic information to dynamically guide network fusion, and adopts a stage-wise training strategy to reduce network complexity.
2) **Proposed a Multiscale Edge Gradient Block (MEGB):** It is designed to adaptively extract multiscale local features, effectively retaining image edge information and significantly enhancing texture details.
3) **Proposed a Semantic Prior Classification Network (SPC-Net):** By establishing global long-range dependencies to extract multi-level information, it introduces class activation weights into feature fusion. This achieves a fully learnable adaptive fusion rule based on binary classification, avoiding the limitations of manually designed rules.

2. Methodology

The overall architecture of the SIDM-Fusion network, illustrated in Figure 1, consists of four main components: the Dual-Branch Layer-Wise Interactive Feature Coding Network (DBL-IFCN), the SPC-Net, the fusion network, and the decoder network. Among them, the decoder network, which reconstructs the fused features into the final image, comprises three 3×3 convolutions and one 1×1 convolution. Its operational principle is straightforward and easy to implement.

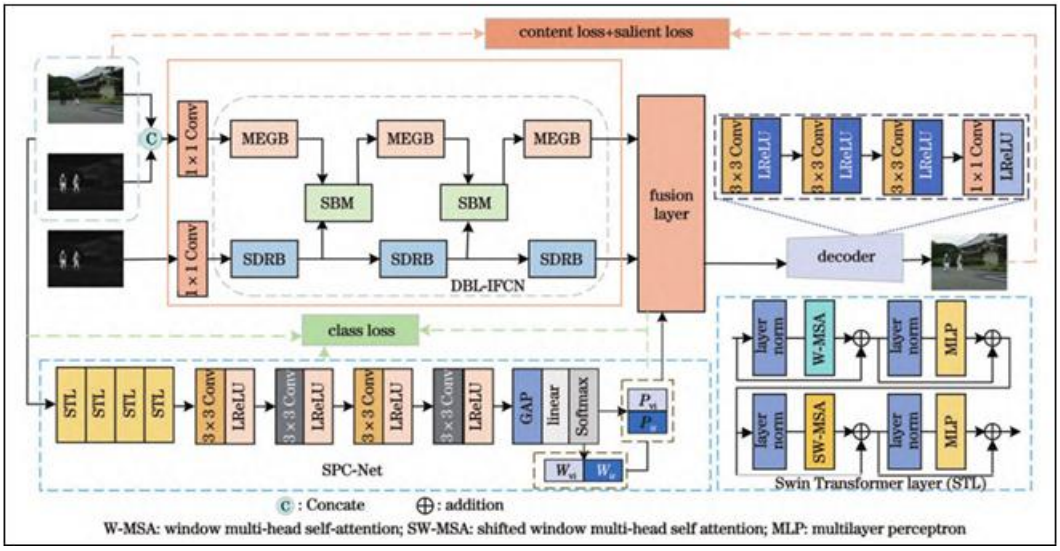

Figure 1: Architecture of our proposed network.

2.1 DBL-IFCN

The DBL-IFCN is composed of the MEGB, the Significantly Dense Residuals Block (SDRB), and the Spatial Bias Module (SBM). It is designed to extract features from both infrared (IR) and visible (VIS) images to acquire texture and semantic information across different modalities.

2.1.1 MEGB

To address the inadequate representation of single-scale feature information, the MEGB is designed utilizing various convolutional kernels. As depicted in Figure 2, it consists of a multi-scale branch and a residual gradient branch. Small convolutional kernels are employed to refine local features and retain critical information, while progressively larger kernels are used to expand the receptive field, facilitating a smooth transition from local details (3×3) to global semantics (7×7). Multi-scale features are fused via concatenation to overcome the limitations of single-scale features and enhance the model's representational capacity. Simultaneously, these multi-scale features interact with features from adjacent branches through element-wise addition, achieving a progressive fusion of local details and global context. Finally, a Sobel operator is incorporated to further preserve edge information. The specific process is formulated as follows:

$$\Phi_T = Conv_{1\times1}\{Conv_{1\times1}(\nabla_{Sobel}\Phi_H) \oplus Conv_{1\times1}[C(\Phi_{1,}\Phi_{3,}\Phi_{5,}\Phi_7)]\} \quad (1)$$

where Φ_H denotes the module input; $Conv_{1\times1}$ represents a 1×1 convolution operation; $C(\cdot)$ indicates concatenation along the channel dimension; ∇_{sobel} represents the Sobel operator; $\oplus$ denotes element-wise addition; $\Phi_n(n \in \{1,3,5,7\})$ represents features of different scales extracted by convolutional kernels of varying sizes; and Φ_T denotes the final extracted feature of the module.

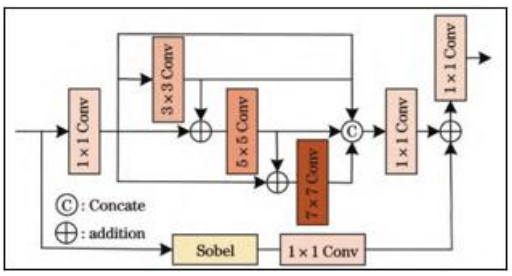

Figure 2: Architecture of MEGB.

2.1.2 SDRB

The SDRB is a residual structure consisting of a dense connection block, a residual block, and a Channel Attention Module (ChAM). It effectively enhances contrast and highlights salient targets, with its detailed structure depicted in Figure 3. The shallow features of the IR image pass through the dense connection block to obtain reused features. These features are then further processed by the ChAM, passing through a 3×3 convolutional kernel, global average pooling, and an activation function to derive the relative feature weights. The specific process is defined as:

$$\Phi_E = C\{\Phi_C, Conv_{3\times3}[Conv_{3\times3}(\Phi_C)], Conv_{3\times3}[Conv_{3\times3}(\Phi_C)]\} \quad (2)$$

$$\Phi_S = Sigmoid\{FC\{GAP[Conv_{1\times1}(\Phi_E)]\}\} \cdot \Phi_E \oplus Conv_{1\times1}(\Phi_C) \quad (3)$$

where $Conv_{3\times3}(\cdot)$ is the 3×3 convolution operation; $GAP(\cdot)$ stands for global average pooling; $FC(\cdot)$ represents a fully connected layer; $Sigmoid(\cdot)$ is the activation function; Φ_C denotes the input shallow features; ϕ_E represents the reused features; and Φ_E is the feature output of the SDRB.

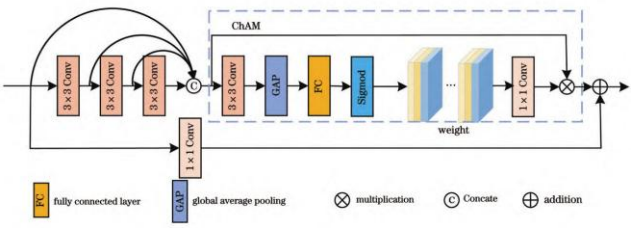

Figure 3: Architecture of SDRB

2.1.3 SBM

The SBM is primarily responsible for the interaction between VIS features and salient IR features. Its structure is shown in Figure 4. It alters the channel number and spatial dimensions of the salient IR features using a 1×1 convolution and max pooling to obtain spatial bias features, effectively capturing global dependencies. Subsequently, the spatial bias features are concatenated with the VIS features along the channel dimension, and a 1×1 convolution is applied to further compress and fuse the information within the feature maps. The SBM mainly involves three parameters: the channel reduction ratio, the max pooling kernel size, and the output channel expansion. Based on relevant literature and experimental results, these parameters are set to 1/4, 2×2, and 16, respectively. The specific operation of the SBM can be expressed as:

$$\Phi_T^{'} = Conv_{1\times1}\{C[\Phi_T, SB(\Phi_S)]\} \quad (4)$$

where $SB(\cdot)$ denotes the spatial bias operation; and $\Phi_T^{'}$ represents the fused features.

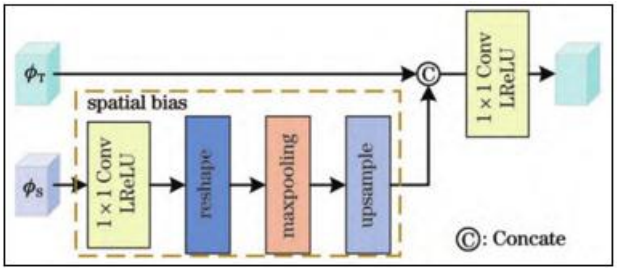

Figure 4: Architecture of SBM

2.2 SPC-Net

To address the limitation of simple convolutions, which focus solely on local texture features and lack long-range semantic dependencies, the SIDM-Fusion network constructs a semantic prior generation network by combining the Swin Transformer Layer (STL) module with convolutions. This allows the network to effectively capture pixel correlation information within windows while enabling cross-window information interaction, thereby focusing on multi-level information within the feature maps. The specific structure is shown in Figure 1. The semantic prior generation network achieves feature extraction through 4 layers of STL modules and 4 layers of 3×3 convolutions, where the stride of the 2nd and 4th convolutional layers is set to 2 to further expand the receptive field. Subsequently, based on the Class Activation Mapping (CAM) mechanism, it utilizes global average pooling and a fully connected layer to learn and acquire class activation weights, outputting the final classification results. This, in turn, provides fusion weights for the fusion layer, realizing a fully learnable fusion rule that eliminates the need for manual design. The specific process can be expressed as:

$$P_c = \sigma\{\sum_m [W_c^m GAP(\Phi_m)]\} \quad (5)$$

where P_c is the probability that the input image belongs to class c, with $c \in \{ir, vi\}$ representing the IR or VIS image; $\sigma(\cdot)$ is the Softmax function; Φ_m represents the features of the m-th channel extracted by the final convolutional layer; and W_c^m is the class activation weight obtained through network training, indicating the significance of this channel's features within the network.

Based on this, after being processed by the fusion layer, the fused features of the m channel can be represented as:

$$\varphi^m = \sum_c \rho(P_c W_c^m)\varphi_c^m \quad (6)$$

where φ_c^m is the feature of the m-th channel for input class c; and $\rho(\cdot)$ denotes the normalization function.

2.3 Loss Function

The SIDM-Fusion network is trained by combining the SPC-Net loss, content loss, and saliency loss. Acting as a classifier, the SPC-Net employs the cross-entropy loss L_{class} to constrain the network training, which is defined as:

$$L_{class} = -P_x \log \sigma(P_C) - (1 - P_x)\log[1 - \sigma(P_C)] \quad (7)$$

The MEGB aims for the fused image to preserve rich texture details while maintaining an optimal intensity distribution. Therefore, a content loss is introduced, which consists of both intensity loss and gradient loss components. It can be formulated as:

$$L_{content} = \frac{1}{HW}\left\| I_f - Max(I_{ir}, I_{vi}) \right\|_1 + \alpha \frac{1}{HW}\left\| \left|\nabla_{Sobel} I_f\right| - Max(\left|\nabla_{Sobel} I_{ir}\right|, \left|\nabla_{Sobel} I_{vi}\right|) \right\|_1 \quad (8)$$

where H and W represent the height and width of the input image, respectively; α is a balancing parameter; I_{ir} and I_{vi} denote the infrared and visible images, respectively; and I_f stands for the fused image.

To preserve the salient targets within the fused image, an intermediate saliency loss is constructed using a target mask, specified as follows:

$$L_{Salient} = \frac{1}{HW}\left\| I_m \cdot I_{ir} - CA(\Psi_{ir}, \Psi_{vi}) \right\|_1 \quad (9)$$

where I_m indicates the target mask; $CA(\cdot)$ represents the channel average value; and Ψ_{ir} and Ψ_{vi} are the deep feature maps of the infrared and visible images, respectively.

Combining the above, the total loss is calculated as:

$$L_{Fusion} = \lambda_1 \cdot L_{Content} + \lambda_2 \cdot L_{Salient} + \lambda_3 \cdot L_{class} \quad (10)$$

where $\lambda_i (i = 1,2,3)$ represent the weight parameters for their respective loss terms.

3. Experiments

3.1 Datasets and Evaluation Metrics

The training set comprises 45 image pairs from TNO, alongside 416 daytime and 354 nighttime pairs from MSRS. The test set includes 20 TNO pairs, plus 42 daytime and 100 nighttime MSRS pairs. All images are standardized to 480 × 640 pixels. We evaluate performance using six positive objective metrics: Mutual Information (MI), Visual Information Fidelity (VIF), Average Gradient (AG), Sum of the Correlations of Differences (SCD), Entropy (EN), and an edge-based metric $Q^{AB/F}$. VIF assesses visual perception, while the rest evaluate spatial information preservation.

3.2 Comparative Experiments and Results Analysis

Tests were conducted on the public TNO and MARS datasets, and the performance was compared against current mainstream methods, including, U2Fusion[12], SeAFusion[13], and PIAFusion[11].

3.2.1 Objective Metric Analysis

Table 1 reports the quantitative results, with the best and second-best values in bold and underlined, respectively. Our model achieves state-of-the-art results in five of the six metrics. PIAFusion yields the highest VIF by explicitly optimizing visual fidelity in its loss function. In contrast, our model prioritizes multi-level semantic preservation, which marginally sacrifices visual fidelity (VIF remains > 0.85, ranking second). Compared to the second-best methods, our model improves MI, AG, SCD, and EN by 14.1%, 21.8%, 2.3%, and 1.9% on TNO, and boosts MI, AG, SCD, and $Q^{AB/F}$ by 22.5%, 9.4%, 7.6%, and 8.3% on MSRS. These gains demonstrate our model's superior comprehensive fusion capability.

3.2.2 Visual Result Analysis

To intuitively demonstrate these advantages, a visual analysis was conducted. Partial visual results are illustrated in Figures 5 and 6, where regions with significant differences are highlighted by red and green bounding boxes. As shown in Figure 5, in night scenes, most methods overemphasize information from a single modality. Specifically, the infrared targets are weakened in DenseFuse[8], RFN-Nest[9], U2Fusion[12], and SeAFusion[13]. Conversely, the fused images generated by FusionGAN[4] and SDNet[10] are heavily biased toward the infrared images, leading to blurred background details. By employing the SBM to facilitate interactions between visible features and salient infrared features, the proposed model effectively integrates complementary information in low-light scenarios.

As shown in Figure 6, in daytime scenes, RFN-Nest and FusionGAN fail to preserve the texture details of the visible images. The fused images from SDNet, DenseFuse, and U2Fusion exhibit blurred edges and low contrast; for example, the silhouette of the person in the right red box remains indistinguishable. While PIAFusion retains texture details relatively well, it introduces minor spectral contamination around salient targets. The proposed model successfully highlights salient targets, ensuring distinct boundaries for low-contrast objects such as pedestrians and warning signs.

Table 1: Performance comparison of different methods across multiple datasets

Dataset	Algorithm	MI	VIF	AG	SCD	EN	$Q^{AB/F}$
TNO	U2Fusion	2.48	0.67	3.48	1.58	6.42	0.32
	SeAFusion	2.40	0.79	3.27	1.71	6.63	0.58
	PIAFusion	3.48	0.88	4.42	1.65	6.89	0.44
	Ours	3.98	0.86	5.50	1.81	7.06	0.62
MSRS	U2Fusion	2.41	0.70	3.85	1.54	6.62	0.32
	SeAFusion	3.69	0.69	2.13	1.49	6.57	0.54
	PIAFusion	3.73	0.93	4.97	1.33	6.80	0.62
	Ours	4.57	0.85	5.43	1.75	6.94	0.68

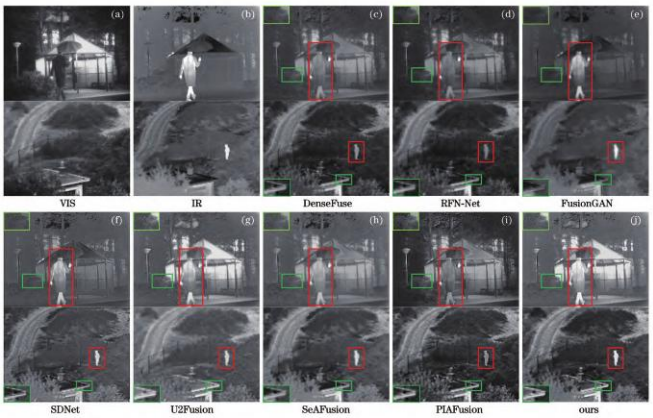

Figure 5: Visual Contrast of Images in the TNO Dataset.

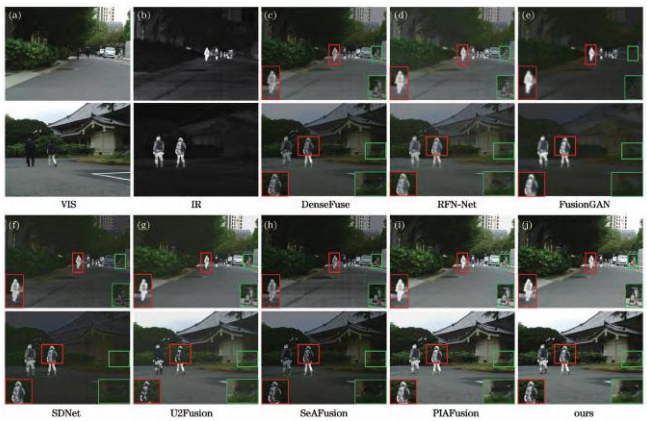

Figure 6: Visual Contrast of Images in the MSRS Dataset

3.3 Ablation Study

To evaluate the impact of different modules on the final reconstruction, we conducted ablation studies on the MSRS dataset. Table 2 presents the quantitative results, where a checkmark (√) denotes the inclusion of a module, a blank

space indicates its omission, and the best results are highlighted in bold. Notably, "MEGB" represents the module without the Sobel operator, while "MEGB*" includes it. The experiments demonstrate that the model achieves optimal fusion performance when all modules are utilized simultaneously.

Figure 7 visualizes these effects. Using only the MEGB yields a relatively smooth scene (Fig. 7c). Adding the Sobel operator to the MEGB (i.e., MEGB*) enhances the image edge information (Fig. 7d). Combining MEGB* with the SBM further improves edge details in the fused image (Fig. 7e), whereas combining MEGB* with the SDRB increases target saliency (Fig. 7f). When MEGB* and the SPC-Net are applied together, background texture details are significantly enhanced (Fig. 7g). The joint application of MEGB*, SDRB, and SBM brightens the overall fused image and improves target contrast (Fig. 7h). Finally, employing all modules jointly effectively mitigates information loss when the visible image is obscured by smoke.

Figure 7: Visual Contrast of Images in the TNO Dataset

4. Conclusions

This paper proposes a Semantic Information Driven Multimodal Image Fusion network (SIDM-Fusion) that dynamically guides the fusion process using semantic information based on the class activation mapping mechanism. Within the network, the Multiscale Edge Gradient Block (MEGB) adaptively extracts multiscale features while preserving edges and enhancing texture details. Concurrently, the Semantic Prior Classification Network (SPC-Net) focuses on multi-level information, achieving fully learnable adaptive fusion rules by introducing class activation weights. Both subjective and objective experiments on public datasets demonstrate that the proposed network outperforms current mainstream methods in terms of fusion quality and downstream semantic segmentation accuracy. However, the model involves a relatively large number of parameters and suffers from noticeable texture loss in visible images under extreme low-light conditions. Future work will integrate a dynamic multiscale selection mechanism and low-light enhancement preprocessing to address these information loss issues and optimize efficiency.

Table 2: Quantitative Evaluation Results of Ablation Study

Group	MEGB	Sobel	SBM	SDRB	SPC-Net	MI	VIF	AG	SCD	EN	$Q^{AB/F}$
1	√					2.24	0.47	3.13	0.71	5.02	0.38
2	√	√				2.37	0.58	3.35	1.27	5.68	0.36
3	√	√	√			2.31	0.64	3.36	1.47	5.73	0.34
4	√	√		√		2.41	0.67	3.18	1.56	5.92	0.46
5	√	√			√	2.58	0.67	2.91	1.57	5.88	0.47
6	√	√	√	√		3.68	0.73	3.25	1.65	6.12	0.553
7	√	√	√	√	√	3.90	0.78	3.37	1.80	6.40	0.64

References

[1] Y. Jie, Y. Xu, X. Li, F. Zhou, J. Lv, and H. Li, "FS-Diff: Semantic guidance and clarity-aware simultaneous multimodal image fusion and super-resolution," *Information Fusion,* vol. 121, p. 103146, 2025.

[2] M. Wang *et al.*, "Task-generalized adaptive cross-domain learning for multimodal image fusion," *IEEE Transactions on Multimedia,* 2026.

[3] M. Zhou *et al.*, "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2024.

[4] 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.

[5] P. Shanmugam and S. A. M. J. Amali, "Dual-discriminator conditional generative adversarial network optimized with hybrid manta ray foraging optimization and volcano eruption algorithm for hyperspectral anomaly detection," *Expert Systems with Applications,* vol. 238, p. 122058, 2024.

[6] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-13, 2021.

[7] J. Liu *et al.*, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5802-5811.

[8] 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, 2018.

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

[10] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion,"

International Journal of Computer Vision, vol. 129, no. 10, pp. 2761-2785, 2021.

[11] 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, pp. 79-92, 2022.

[12] 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, 2020.

[13] 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.

[14] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Detfusion: A detection-driven infrared and visible image fusion network," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 4003-4011.

[15] S. Kalamkar, "Multimodal image fusion: A systematic review," *Decision Analytics Journal,* vol. 9, p. 100327, 2023.