Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 3, 2026

Infrared and Visible Image Fusion Algorithm Based on Cross-Modal Attention Mechanism

Zhiyuan Wang

Zone 10, Headquarters Base of Fengtai District, Beijing, China

Abstract: As a critical branch of multi-modal image processing, infrared and visible image fusion boasts high application value in intelligent security and has drawn widespread global research attention. Enhancing model feature extraction is a core scientific challenge in this field. This paper presents a dual-branch infrared and visible image fusion algorithm based on feature decomposition. In fusion tasks, shared features characterize global information while private features focus on local details. To boost feature representation, we design a feature decomposition module that splits shallow features into shared and private components: a coarse-grained branch with medium-to-large receptive fields handles global shared features, and a fine-grained branch with small receptive fields extracts local private features. Parallel dual-branch processing of decomposed features enables precise data structure mining, reduces redundancy, and efficiently captures key information. Experiments on four mainstream public datasets validate that the proposed algorithm surpasses state-of-the-art methods in information extraction, detail preservation and fusion performance.

Keywords: Image fusion, attention mechanism, Swin Transformer.

1. Introduction

With the rapid development of information technology, image fusion technology plays an increasingly critical role in multiple fields such as security monitoring, target detection, remote sensing imaging and night vision. Among them, infrared and visible image fusion has gained widespread attention from academia and industry by virtue of its modal complementarity advantages. This technology can effectively integrate the thermal radiation target information of infrared images and the texture detail features of visible images, generating fused images with richer information that are more suitable for human eye observation and machine recognition. However, achieving efficient and robust infrared and visible image fusion still faces numerous technical bottlenecks. One of the core difficulties is how to accurately extract and efficiently fuse complementary features from heterogeneous source images.

Traditional image fusion methods are mostly based on simple pixel-level or shallow feature-level superposition, which makes it difficult to fully mine the deep semantic information of images, resulting in limited generalization ability and fusion accuracy. As a classic deep learning architecture, Convolutional Neural Networks (CNNs) show powerful advantages in local image feature extraction. Nevertheless, existing CNN-based fusion models often overly focus on local feature learning, ignoring the long-range dependencies between features and restricting the global feature representation capability. The emergence of Transformer models provides a novel solution to this problem. Relying on the self-attention mechanism, Transformer can effectively capture long-range dependencies of sequential data and has achieved breakthrough progress in the field of natural language processing; subsequently, Transformer was introduced into computer vision tasks and also exhibited excellent performance. As an improved visual variant of Transformer, Swin Transformer further enhances the model's global and local image feature extraction capability while reducing computational overhead by introducing a hierarchical sliding window attention mechanism.

In spite of this, existing Transformer-based image fusion models still have obvious shortcomings: such models mostly focus on attention calculation within a single modal domain, only performing feature extraction for images of the same modality, ignoring the information interaction and complementarity between infrared and visible modalities, thereby limiting the improvement of fused image quality. To address the above problems, this paper proposes an infrared and visible image fusion method based on Swin Transformer. This method takes Swin Transformer as the backbone network, and directly generates fusion weights to act on the pixels of original images, minimizing pixel distortion that may occur during network reconstruction. In addition, to strengthen the model's cross-modal information interaction capability for heterogeneous source image features, a dedicated cross-modal attention mechanism is designed to promote the deep fusion of infrared and visible images at the feature level, ultimately generating fused images with more comprehensive information, clearer details and better visual effects.

2. Related Work

Over recent decades, image fusion techniques have evolved into two main categories: traditional methods and deep learning-based approaches [1]. Traditional methods (e.g., sparse representation [2-4], multi-scale transformation [5-7]) rely on handcrafted features and fusion rules, resulting in limited generalization and robustness. Benefiting from deep learning, infrared and visible image fusion methods have overcome these drawbacks and become the mainstream with stronger adaptability. This section reviews typical deep learning fusion algorithms and attention-driven fusion strategies.

2.1. Deep Learning-Based Image Fusion Algorithms

Early deep learning fusion methods relied on manual fusion

rules [8], which are inefficient and prone to subjective bias. In 2017, Liu et al. [9] first introduced CNNs into image fusion, enabling joint learning of feature extraction and fusion rules; Zhang et al. [10] proposed the generic IFCNN framework in the same year. Xu et al. [11] further developed U2Fusion (2022) for adaptive multi-task fusion.

Autoencoder (AE)-based architectures have been widely adopted: Li et al. proposed DenseFuse (2018) [12] and RFN-Nest (2021) [13], abandoning manual rules for end-to-end feature fusion. Hong et al. [14] enhanced feature representation via multi-branch encoders in 2022.

Generative Adversarial Networks (GANs) were applied by Ma et al. with FusionGAN (2019) [15], followed by DDcGAN (2020) [16] and UIFGAN (2023) [17] to boost cross-modal information learning. Recently, Transformer-based models (e.g., IFT [18]) have emerged to capture long-range feature dependencies, outperforming conventional CNNs in global modeling.

2.2. Attention Mechanism-Based Fusion Algorithms

Attention mechanisms, simulating human visual selectivity, have become pivotal for image fusion. Li et al. integrated spatial-channel attention into NestFuse (2020) [19]; Wang et al. designed non-local attention in Rest2Fusion (2021) [20]; Tang et al. combined channel-spatial attention with residual learning in DATFuse (2023) [21] to preserve fine details.

Swin Transformer [22], with hierarchical window attention, balances efficiency and performance, spawning SwinFuse [23] and SwinFusion [24]. However, most models focus on single-modal attention and neglect cross-modal interaction. Ma et al. [24] addressed this gap with cross-modal self-attention, which inspires the cross-modal attention mechanism designed in this work to optimize feature interaction and fusion quality.

3. Algorithm Design

The Swin Transformer delivers superior performance in various vision tasks and serves as an effective backbone for end-to-end fusion networks, alleviating the feature locality issue in infrared and visible image fusion. On this basis, this work further explores performance improvements via cross-modal feature extraction and adaptive fusion strategies.

3.1. Overall Network Architecture Based on Swin Transformer

The overall framework of the proposed cross-modal attention-driven infrared and visible image fusion algorithm is illustrated in Fig. 1. The infrared and visible images are fed into a Swin Transformer-based backbone network to learn adaptive fusion weights, which are directly computed with raw image pixels to generate the final fused result. This weight-based fusion scheme preserves original pixel information to the maximum extent and avoids distortion caused by feature extraction and reconstruction.

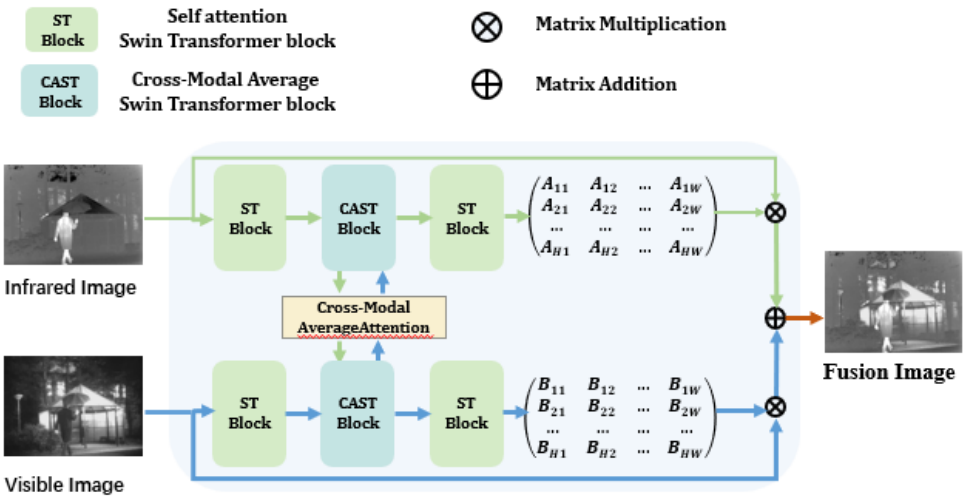

Figure 1. Overall Framework Structure of the Infrared and Visible Image Fusion Algorithm Based on Cross-Modal Attention Mechanism

The backbone consists of two Swin Transformer blocks (STBlock) based on self-attention and one Cross-Modal Average Swin Transformer block (CASTBlock) based on cross-modal average attention, with nomenclature corresponding to their attention mechanisms. Each STBlock contains four standard Swin Transformer layers (STL), which compute intra-modal features using self-attention. Each CASTBlock contains two Cross-Modal Average Swin Transformer layers (CASTL), which implement joint feature learning across heterogeneous modalities via Cross-Modal Average Attention (CAAttention). Infrared and visible images enter the model through independent channels, undergo multi-level feature extraction, and integrate cross-domain information to generate the final fusion weights.

3.2. Design of Cross-Modal Attention Mechanism

The Swin Transformer realizes intra-window and inter-window information interaction via a sliding window mechanism, but it lacks efficient cross-modal feature integration. To overcome this limitation, a cross-modal average attention mechanism is proposed, which enhances the perception of heterogeneous modal features on top of self-attention and promotes cross-modal information complementarity. The difference between self-attention and the proposed cross-modal average attention is visualized in Fig. 2.

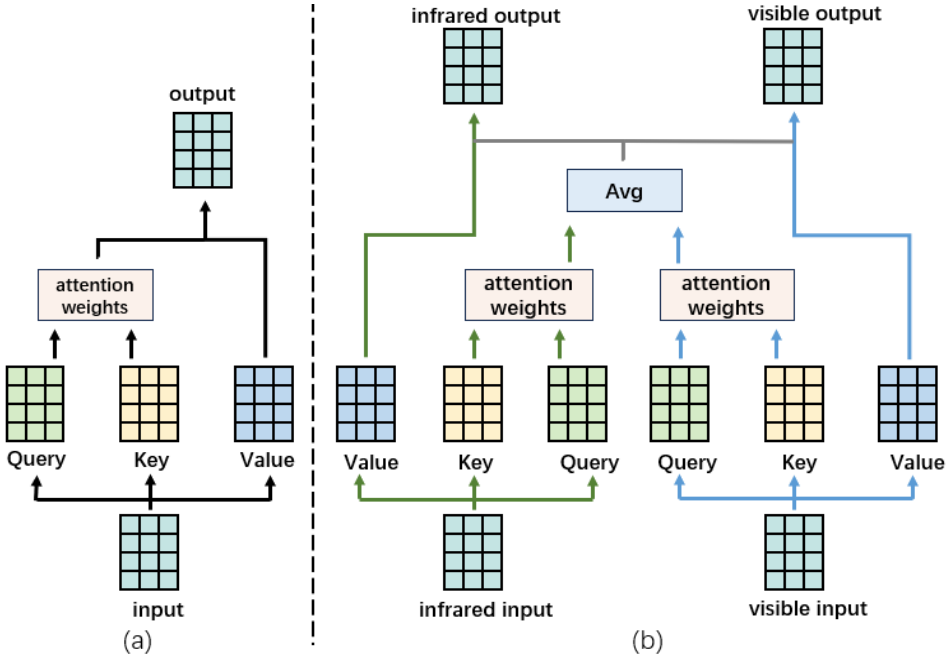

Figure 2. Self-attention mechanism (left) and cross-modal average attention mechanism (right)

Taking the infrared image as an example, the cross-modal average attention mechanism is formulated as follows:

$$\{Q_{ir}, K_{ir}, V_{ir}\} = \{X_{ir}W_{ir}^{Q}, X_{ir}W_{ir}^{K}, X_{ir}W_{ir}^{V}\} \quad (1)$$

$$\{Q_{vi}, K_{vi}, V_{vi}\} = \{X_{vi}W_{vi}^{Q}, X_{vi}W_{vi}^{K}, X_{vi}W_{vi}^{V}\} \quad (2)$$

$$CAAttention(Q_{ir}, K_{ir}, Q_{vi}, K_{vi}, V_{ir}) = \text{softmax}\left(\frac{Q_{ir}K_{ir}^{T} + Q_{vi}K_{vi}^{T}}{2\sqrt{d_k}} + B\right)V_{ir} \quad (3)$$

Where X_{ir} and X_{vi} denote the window features of infrared and visible images, respectively; Q, K and V represent the query, key, and value vectors in attention computation; d_k is the dimension of the key vector, and B denotes the learnable positional embedding. For infrared features, CAAttention averages the attention weights of infrared and visible modalities before applying them to the infrared value vector, realizing cross-modal information interaction. Replacing self-attention in STL with CAAttention yields CASTL. Two consecutive CASTL layers are depicted in Fig.3, and the infrared feature computation is expressed as:

$$\hat{z}_{ir}^{l} = W - MCAA\left(LN\left(z_{ir}^{l-1}\right)\right) + z_{ir}^{l-1} \quad (4)$$

$$z_{ir}^{l} = MLP\left(LN\left(\hat{z}_{ir}^{l}\right)\right) + \hat{z}_{ir}^{l} \quad (5)$$

$$\hat{z}_{ir}^{l+1} = SW - MCAA\left(LN\left(z_{ir}^{l}\right)\right) + z_{ir}^{l} \quad (6)$$

$$z_{ir}^{l+1} = MLP\left(LN\left(\hat{z}_{ir}^{l+1}\right)\right) + \hat{z}_{ir}^{l+1} \quad (7)$$

Where z_{ir}^{l} denotes the output infrared feature of the l-th CASTL layer; $W-MCAA$ refers to window-based multi-head cross-modal average attention, and $SW-MCAA$ denotes shifted window-based multi-head cross-modal average attention.

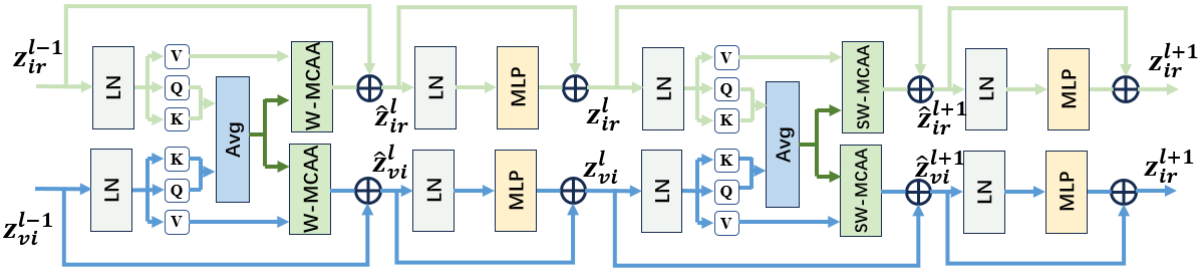

Figure 3. Two consecutive CASTL layers

3.3. Loss Function

The proposed model is an end-to-end architecture that takes paired infrared and visible images as input and directly outputs the fused image. To constrain network training precisely, a composite loss function is designed, consisting of Structural Similarity (SSIM) loss [71], intensity loss, and texture loss. This multi-component loss optimizes fusion quality from structural integrity, detail preservation, and illumination intensity perspectives.

SSIM measures image similarity by evaluating luminance, contrast, and structure, with a range of [-1, 1]; higher values indicate greater similarity. We adopt SSIM as the structural loss, defined as:

$$L_{ssim} = w_1 \cdot \left(1 - ssim(I_f, I_{ir})\right) + w_2 \cdot \left(1 - ssim(I_f, I_{vi})\right) \quad (8)$$

Where I_f, I_{ir} and I_{vi} represent the fused image, infrared input, and visible input, respectively. The weights w_1 and w_2 are both set to 0.5 to balance the retention of infrared and visible information.

Texture loss is designed to preserve fine details from source images, computed using the Sobel gradient operator and L1 norm:

$$L_{text} = \frac{1}{HW}\left\| |\nabla I_f| - \max(|\nabla I_{ir}|, |\nabla I_{vi}|) \right\| \quad (9)$$

Where ∇ denotes the gradient operation, $|\cdot|$ denotes absolute value, $\max(\cdot)$ denotes pixel-wise maximum selection, and $\|\cdot\|$ denotes the L1 norm.

Intensity loss optimizes the illumination distribution of fused images for better visual quality and target detectability, formulated as:

$$L_{int} = \frac{1}{HW}\left\| I_f - \mathrm{M}(I_{ir}, I_{vi}) \right\| \quad (10)$$

Where $M(\cdot)$ denotes the pixel-wise maximum operation between infrared and visible images.

The total loss function combines the three complementary terms as follows:

$$L_{sum} = L_{ssim} + L_{text} + L_{int} \quad (11)$$

4. Experimental Results and Analysis

This section evaluates the proposed algorithm on four widely used public datasets: TNO, MSRS, M3FD, and RoadScene. All experiments are conducted on a workstation equipped with an NVIDIA Tesla T4 GPU, implemented in PyTorch 1.11.0 and Python 3.7. Experiments include parameter settings, ablation studies, and comparative evaluations against state-of-the-art methods.

4.1. Experimental Parameter Settings

A total of 1,083 infrared-visible image pairs from the MSRS dataset are used for training, with the remaining 361 pairs for testing. Additional test sets include 42 pairs from TNO, 300 pairs from M3FD, and 221 pairs from RoadScene to verify generalization. During training, 128×128 image patches are randomly cropped; the training epoch is set to 300, batch size to 16, and all loss weights to 1. The ADAM optimizer is adopted with an initial learning rate of 0.0002 and exponential decay. The end-to-end model directly outputs fused results given paired input images.

4.2. Ablation Experiments

Ablation studies are conducted to determine the optimal layer configuration of STBlocks and CASTBlocks. Three architectures (6-6-6, 4-2-4, 4-6-2) are compared on the TNO dataset, with results listed in Table 1. After comprehensive analysis, the 4-6-2 (STBlock-CASTBlock-STBlock) configuration is adopted for its balanced performance.

Table 1. Objective metric results of layer structure design ablation on the TNO dataset (optimal values are shown in bold)

	SD	MI	VIF	SCD	EN	Qabf	SF
6, 6, 6	43.266	2.600	0.647	1.799	6.570	0.460	11.235
4, 6, 2	41.285	4.388	0.862	1.668	6.864	0.599	11.189
4, 2, 4	44.352	3.816	0.796	1.627	6.914	0.583	11.398

To validate the effectiveness of cross-modal average attention, a comparison between self-attention and CAAttention is performed (Table 2). CAAttention significantly improves MI, VIF, and Qabf—metrics that reflect information retention from source images. Although slightly inferior in SD and SCD, the overall performance of CAAttention is distinctly superior.

Table 2. Objective metric results of attention mechanism ablation on the TNO dataset (optimal values are shown in bold)

	SD	MI	VIF	SCD	EN	Qabf	SF
Self-attention	42.929	2.671	0.654	1.809	6.549	0.462	11.170
CAAttention	41.285	4.388	0.862	1.668	6.864	0.599	11.189

4.3. Subjective and Objective Evaluation

The proposed method is compared with 10 state-of-the-art fusion algorithms: DenseFuse [25], FusionGAN [26], NestFuse [27], IFCNN [28], SDNET [29], Res2Fusion [30], U2Fusion [31], SwinFusion [32], DATFuse [33], and DIVFusion [34]. Seven quantitative metrics are used for objective assessment, paired with subjective visual comparison.

4.3.1. Results on the TNO Dataset

A field road scene from TNO is selected for subjective evaluation (Fig. 4). The pedestrian (red box) is only visible in the infrared image, while the fence (blue box) is clear in the visible image. The proposed method preserves both the salient pedestrian and sharp fence details, outperforming comparative methods. As shown in Table 3, our method ranks first in MI, Qabf, and VIF, with strong overall balance across all metrics.

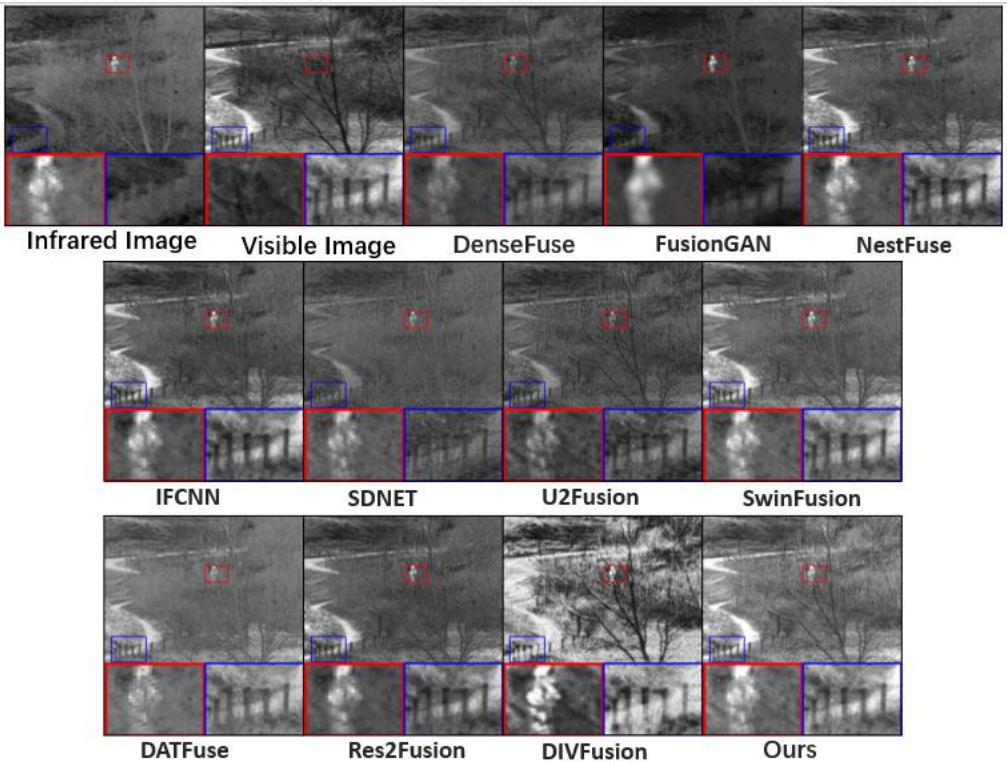

Figure 4. Subjective comparison of fusion results on the TNO dataset

Table 3. Objective metric comparison results on the TNO dataset (optimal values are shown in bold)

	SD	MI	VIF	SCD	EN	Qabf	SF
DenseFuse	26.033	2.096	0.561	1.617	6.281	0.335	6.478
FusionGAN	30.936	1.912	0.168	1.232	6.569	0.133	5.580
NestFuse	41.693	2.635	0.674	1.593	6.744	0.342	13.109
IFCNN	33.585	2.415	0.645	1.685	6.741	0.508	12.018
SDNET	26.903	2.231	0.537	1.455	6.352	0.428	9.430
U2Fusion	25.949	1.873	0.555	1.586	6.423	0.425	8.338
SwinFusion	39.198	3.492	0.752	1.662	6.887	0.527	10.275
DATFuse	27.576	3.132	0.683	1.496	6.453	0.558	9.606
Res2Fusion	39.918	3.496	0.819	1.649	6.960	0.423	9.746
DIVFusion	53.879	2.222	0.625	1.494	7.593	0.312	13.463
Ours	41.285	4.388	0.862	1.668	6.864	0.599	11.189

4.3.2. Results on the MSRS Dataset

An urban intersection scene from MSRS is tested (Fig.5), containing road signs (visible-dominant) and pedestrians (infrared-dominant). Our method clearly retains both text details and complete pedestrian contours, realizing effective cross-modal fusion. Table 4 shows our method achieves the best MI score and top-3 performance in most other metrics.

Table 4. Objective metric comparison results on the MSRS dataset (optimal values are shown in bold)

	SD	MI	VIF	SCD	EN	Qabf	SF
DenseFuse	26.160	2.030	0.429	1.170	4.721	0.289	6.335
FusionGAN	17.070	1.892	0.443	0.983	5.432	0.140	4.351
NestFuse	40.641	4.183	0.973	1.588	6.582	0.663	10.121
IFCNN	35.322	2.850	0.807	1.521	6.439	0.613	12.047
SDNET	14.997	1.693	0.457	0.875	5.026	0.316	7.473
U2Fusion	18.869	1.958	0.474	1.006	4.953	0.315	6.712
SwinFusion	42.878	4.584	0.993	1.675	6.641	0.658	11.035
DATFuse	36.476	3.897	0.906	1.410	6.480	0.640	10.927
Res2Fusion	39.824	4.230	0.922	1.437	6.517	0.656	10.076
DIVFusion	53.411	2.394	0.784	1.197	7.517	0.341	12.598
Ours	44.218	5.363	0.970	1.628	6.550	0.625	11.136

Figure 5. Subjective comparison of fusion results on the MSRS dataset

4.3.3. Results on the M3FD Dataset

A smoke-obscured forest scene from M3FD is evaluated (Fig. 6). Visible images fail to capture pedestrians under smoke, while infrared images highlight thermal targets. Our method prominently enhances pedestrian targets and restores stone texture details. As listed in Table 5, our method achieves the highest MI and VIF scores, demonstrating superior information fusion.

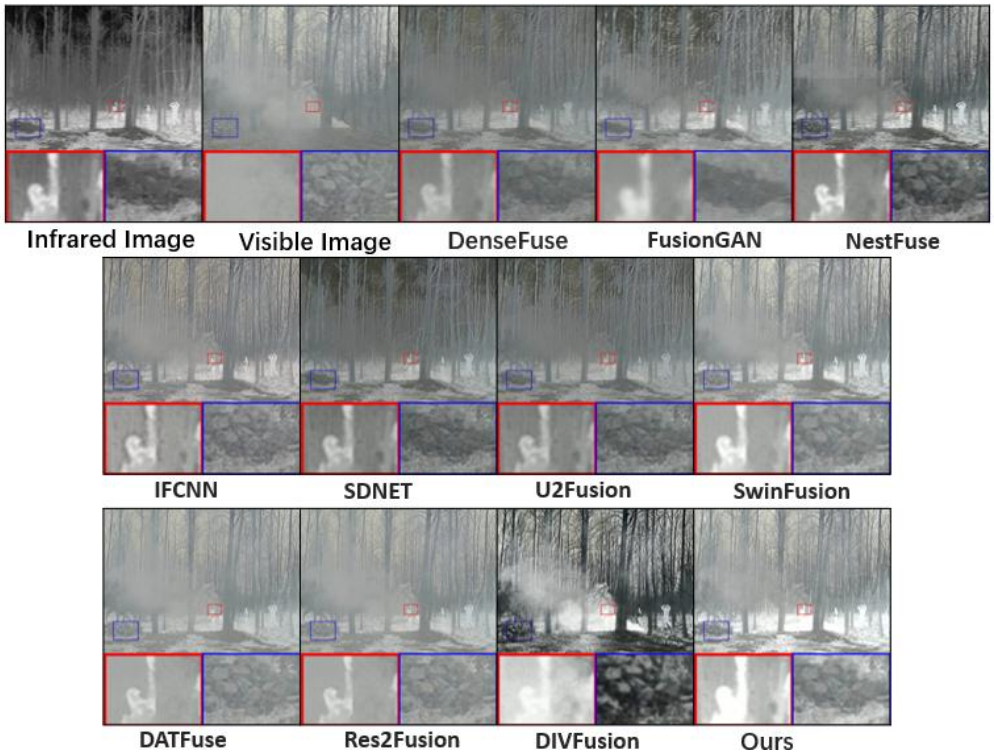

Figure 6. Subjective comparison of fusion results on the M3FD dataset

4.3.4. Results on the RoadScene Dataset

A typical night road scene from RoadScene is tested (Fig. 7). Visible images suffer from light pollution and blurry distant contours, while infrared images capture clear edges but lack signal details. Our method effectively integrates complementary features, preserving both background layering and clear sign text. Although not optimal in all metrics, the gaps to top performers are negligible (Table 6).

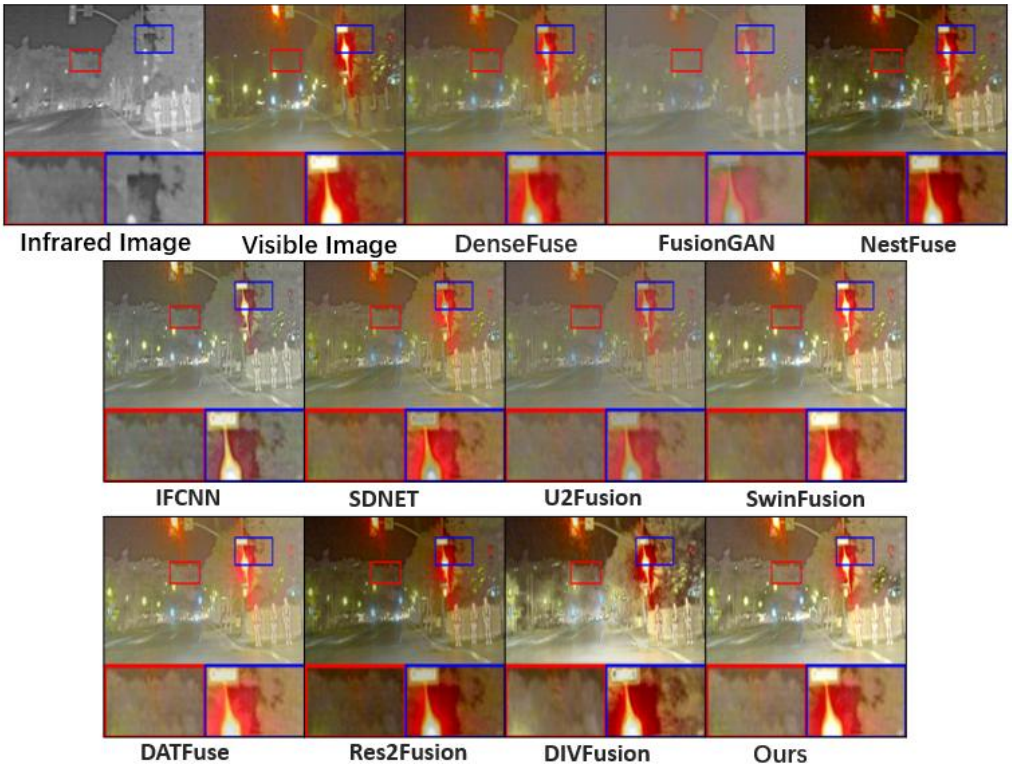

Figure 7. Subjective comparison of fusion results on the RoadScene dataset

Table 5. Objective metric comparison results on the M3FD dataset (optimal values are shown in bold)

	SD	MI	VIF	SCD	EN	Qabf	SF
DenseFuse	26.045	2.866	0.593	1.544	6.433	0.357	7.546
FusionGAN	26.069	2.707	0.121	0.953	6.448	0.116	6.577
NestFuse	35.34	3.827	0.773	1.468	6.849	0.545	11.516
IFCNN	30.274	2.956	0.673	1.555	6.697	0.607	14.593
SDNET	31.237	3.203	0.551	1.42	6.66	0.515	12.105
U2Fusion	28.152	2.753	0.63	1.549	6.629	0.532	10.408
SwinFusion	34.784	4.316	0.766	1.476	6.769	0.607	13.244
DATFuse	26.308	4.130	0.644	1.287	6.402	0.494	10.467
Res2Fusion	35.088	3.970	0.700	1.296	6.534	0.511	12.356
DIVFusion	52.731	2.750	0.714	1.53	7.552	0.431	15.483
Ours	39.462	4.837	0.894	1.485	6.912	0.661	14.269

Table 6. Objective metric comparison results on the RoadScene dataset (optimal values are shown in bold)

	SD	MI	VIF	SCD	EN	Qabf	SF
DenseFuse	32.223	2.911	0.559	1.365	6.817	0.381	8.471
FusionGAN	32.717	2.189	0.142	0.767	6.792	0.147	7.041
NestFuse	56.465	3.838	0.586	1.257	7.275	0.374	15.388
IFCNN	39.199	2.915	0.582	1.441	7.124	0.537	15.276
SDNET	37.675	3.230	0.570	1.190	7.058	0.505	12.808
U2Fusion	31.554	2.645	0.529	1.208	6.831	0.473	11.868
SwinFusion	44.053	3.638	0.647	1.516	6.968	0.465	12.169
DATFuse	31.748	3.668	0.597	1.146	6.720	0.472	11.359
Res2Fusion	47.897	3.397	0.651	1.646	7.338	0.494	12.971
DIVFusion	54.189	2.902	0.572	1.524	7.539	0.342	13.296
Ours	47.658	3.050	0.636	1.603	6.800	0.525	13.703

5. Conclusion

This paper proposes an infrared and visible image fusion network based on cross-modal average attention with Swin Transformer as the backbone. The cross-modal attention mechanism enhances heterogeneous feature interaction via weighted average computation, while direct pixel-level

weight fusion minimizes reconstruction distortion. The composite loss function further optimizes structural integrity, texture clarity, and intensity distribution. Ablation experiments verify the effectiveness of the proposed architecture and attention mechanism. Extensive tests on four public datasets demonstrate that our method outperforms most state-of-the-art algorithms in information retention, detail preservation, and visual quality.

References

[1] Xu H, Ma J, Le Z, et al. FusionDN: A Unified Densely Connected Network for Image Fusion [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 12484-12491.

[2] Zhang Q, Fu Y, Li H, et al. Dictionary learning method for joint sparse representation-based image fusion [J]. Optical Engineering, 2013, 52(5): 057006.

[3] Li H, Wu X J, Kittler J. MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion [J]. IEEE Transactions on Image Processing, 2020, 29: 4733-4746.

[4] Bin Y, Chao Y, Guoyu H. Efficient image fusion with approximate sparse representation [J]. International Journal of Wavelets, Multiresolution and Information Processing, 2016, 14(04): 1650024.

[5] Song C, Gao X, Qiao Y L, et al. Infrared and visible image fusion based on oversampled graph filter banks [J]. Journal of Electronic Imaging, 2020, 29(2): 023016.

[6] He K, Zhou D, Zhang X, et al. Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain [J]. Journal of Applied Remote Sensing, 2017, 11(1): 015011.

[7] Zhi-She W, Feng-Bao Y, Zhi-Hao P, et al. Multi-sensor image enhanced fusion algorithm based on NSST and top-hat transformation [J]. Optik, 2015, 126(23): 4184-4190.

[8] Lu M, Jiang M, Kong J, et al. LDRepFM: A Real-Time End-to-End Visible and Infrared Image Fusion Model Based on Layer Decomposition and Re-Parameterization [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.

[9] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network [J]. Information Fusion, 2017, 36: 191-207.

[10] Zhang Y, Liu Y, Sun P, et al. IFCNN: A general image fusion framework based on convolutional neural network [J]. Information Fusion, 2020, 54: 99-118.

[11] Xu H, Ma J, Jiang J, et al. U2Fusion: A Unified Unsupervised Image Fusion Network [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[12] Li H, Wu X J. DenseFuse: A Fusion Approach to Infrared and Visible Images [J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.

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

[14] Hong Y, Wu X J, Xu T. MEFuse: end-to-end infrared and visible image fusion method based on multibranch encoder [J]. Journal of Electronic Imaging, 2022, 31(3): 033043.

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

[16] Ma J, Xu H, Jiang J, et al. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion [J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995.

[17] Le Z, Huang J, Xu H, et al. UIFGAN: An unsupervised continual-learning generative adversarial network for unified image fusion [J]. Information Fusion, 2022, 88: 305-318.

[18] Vs V, Jose Valanarasu J M, Oza P, et al. Image Fusion Transformer [C]//2022 IEEE International Conference on Image Processing (ICIP). 2022: 3566-3570.

[19] Li H, Wu X J, Durrani T. NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models [J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656.

[20] Wang Z, Wu Y, Wang J, et al. Res2Fusion: Infrared and Visible Image Fusion Based on Dense Res2net and Double Nonlocal Attention Models [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[21] Tang W, He F, Liu Y, et al. DATFuse: Infrared and Visible Image Fusion via Dual Attention Transformer [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3159-3172.

[22] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[23] Wang Z, Chen Y, Shao W, et al. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[24] Ma J, Tang L, Fan F, et al. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer [J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.

[25] Li H, Wu X J. DenseFuse: A Fusion Approach to Infrared and Visible Images [J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.

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

[27] Li H, Wu X J, Durrani T. NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models [J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656.

[28] Zhang Y, Liu Y, Sun P, et al. IFCNN: A general image fusion framework based on convolutional neural network [J]. Information Fusion, 2020, 54: 99-118.

[29] Zhang H, Ma J. SDNet: A Versatile Squeeze-and-Decomposition Network for Real-Time Image Fusion [J]. International Journal of Computer Vision, 2021, 129(10): 2761-2785.

[30] Wang Z, Wu Y, Wang J, et al. Res2Fusion: Infrared and Visible Image Fusion Based on Dense Res2net and Double Nonlocal Attention Models [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[31] Xu H, Ma J, Jiang J, et al. U2Fusion: A Unified Unsupervised Image Fusion Network [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[32] Ma J, Tang L, Fan F, et al. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer [J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.

[33] Tang W, He F, Liu Y, et al. DATFuse: Infrared and Visible Image Fusion via Dual Attention Transformer [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3159-3172.

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