Computers and Artificial Intelligence；Vol.2, No.1; 2025
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.25v2n1.2531
Published by: Zeus Press

Multi-Window Detail Enhancement Based Infrared and Visible Image Fusion

Long Wang* and Xinbo Wang

College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China

**Corresponding author: Long Wang.*

Abstract

The goal of infrared and visible image fusion is to combine complementary information from infrared and visible images of the same scene to generate a high-quality composite image that integrates the advantages of both modalities. Although many current fusion methods achieve satisfactory results, they still suffer from limitations such as insufficient feature resolution extraction from infrared images and inadequate texture information extraction from visible images. This paper proposes a novel fusion method designed to enhance the resolution, detail preservation, and visual consistency of the fused image. The method integrates multi-window detail enhancement with multi-layer residual connections, employing a detail selector and a global feature extractor to separately capture high-frequency and low-frequency features from the infrared and visible images. Experimental results demonstrate that, compared to existing approaches, the proposed method achieves superior fusion quality and better preservation of image details, providing higher-quality data for subsequent image processing tasks.

Keywords

infrared and visible image fusion, multi-window attention, deep learning

1. Introduction

In the field of modern information technology, image fusion has gradually become a key technology for enhancing the value and applicability of images (Park et al., 2003). Infrared and visible image fusion, as an important branch of image fusion technology, has attracted considerable attention due to its unique advantages in multimodal information integration (Toet, 1989). Visible images, with their rich color and detailed information, are widely used in many fields, while infrared images, owing to their imaging capabilities in low-light or special conditions, have demonstrated unique value in military and industrial applications (Ma, Ma, et al., 2019). To address these issues, various infrared and visible image fusion methods have been proposed to combine the strengths of both modalities to generate images with richer information. These methods show great potential in enhancing the visual quality, usability, and application scope of fused images (Pizurica et al., 2003). For example, Zhang et al. (2020) proposed a convolutional neural network (CNN)-based fusion framework that effectively fuses infrared and visible images and achieved excellent performance on multiple datasets. Zhao et al. proposed MetaFusion (Zhao et al., 2023), which further introduced a meta-feature embedding strategy guided by specific tasks to improve cross-modal semantic alignment. Generative Adversarial Networks (GANs) have also been leveraged to enhance the

realism and visual quality of fused images. In 2019, Ma et al. proposed FusionGAN (Ma, Yu, et al., 2019), which adopts a generator–discriminator architecture to improve the naturalness of the fusion results. However, these deep learning-based methods also face challenges such as high model complexity and the need for extensive computational resources and training data. Besides deep learning approaches, traditional image fusion techniques continue to evolve. For instance, Liu et al. (2015) proposes a general image fusion framework based on multi-scale transform and sparse representation, which aims to effectively extract and fuse the key information of multi-source images by combining the advantages of multi-scale analysis and sparse representation, so as to improve the quality and performance of fusion images. Zhou et al. (2023) proposed to achieve low latency and low power motion recognition by performing computation directly within the sensor.

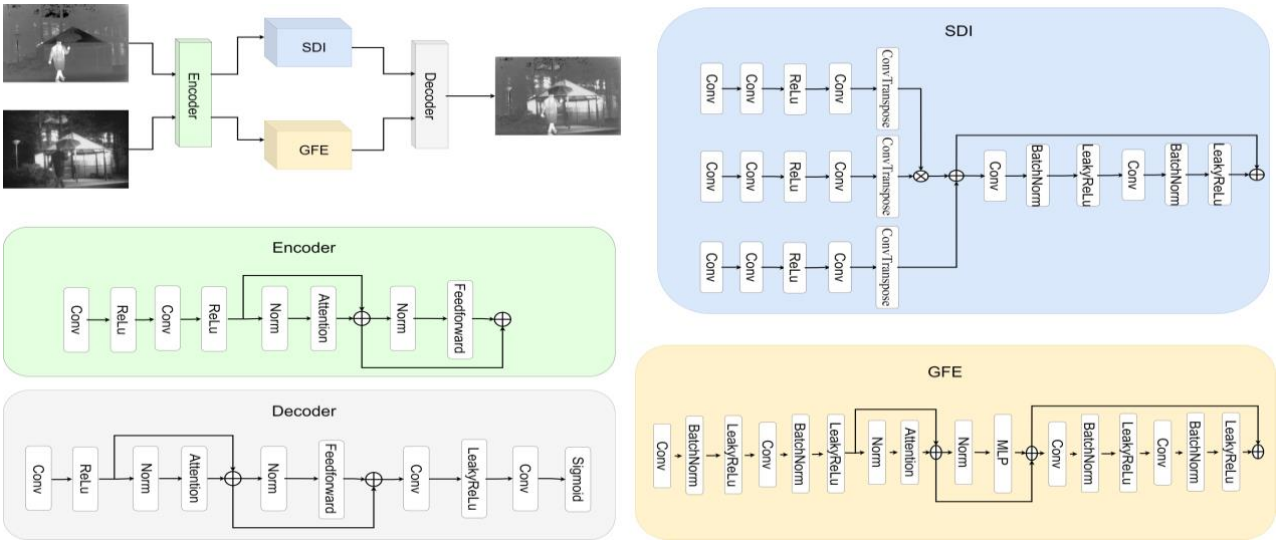
Furthermore, researchers have conducted extensive evaluations and comparisons of image fusion algorithms to identify the most suitable methods for specific applications. These assessments help clarify the strengths and weaknesses of different approaches and provide guidance for future research directions.

2. Method

This paper proposes a model composed of four core components: an encoder, a decoder, a local feature extractor—Selective Detail Integrator (SDI), and a global feature extractor—Global Feature Extractor (GFE). This architecture aims to achieve efficient fusion of infrared and visible images by combining the strengths of Transformer and traditional convolutional neural networks, as illustrated in Fig. 1. The encoder employs Transformer blocks inspired by the Vision Transformer (ViT) architecture to perform feature extraction and encoding. The self-attention mechanism of the Transformer effectively captures long-range dependencies in the images, which is particularly crucial for the fusion of information from infrared and visible modalities. Through multiple layers of Transformer blocks, the encoder progressively extracts multi-scale features from both infrared and visible images, providing rich feature representations for subsequent fusion steps.Following feature extraction, the local and global feature extractors undertake distinct functions.

The SDI module, based on multi-window attention and residual connections, performs multi-scale detail extraction and selective enhancement on the input features, thereby supplying high-quality detail information for subsequent super-resolution reconstruction. In contrast, the GFE focuses on extracting low-frequency information, preserving the global structure and large-scale visual context of the images. While convolutional neural networks excel at capturing local details, Transformers effectively model global features, offering greater flexibility and accuracy in information extraction for image fusion. Through the cooperative operation of these two extractors, the model efficiently processes both fine details and global information, ensuring that the fusion results are both detailed and globally consistent. To guarantee accurate feature representation and fusion across different modalities, a combination of loss functions is adopted, including gradient loss, mean squared error loss, and structural similarity loss.

Figure 1: Overall network diagram



3. Experiment

3.1 Datasets

We trained our model on the training set of the MSRS (Multispectral Remote Sensing) dataset. Each image pair in the training set consists of an infrared image and its corresponding visible image, covering a variety of environments and scenes. By utilizing this dataset, we enable the model to perform image classification tasks under diverse conditions, thereby enhancing its generalization capability and classification accuracy. We selected the TNO (Toet, 2017) and RoadScene (Xu et al., 2022) datasets to conduct a series of experimental evaluations.

The TNO dataset contains multi-spectral remote sensing image pairs, covering a variety of different ground object types and complex environments, which is highly challenging. The RoadScence dataset mainly contains image data of road scenes, and focuses on testing the image fusion performance of the model under complex backgrounds, especially for applications in the field of traffic monitoring and autonomous driving. In this experiment, we evaluate the performance of the model in terms of image quality, fusion effect, and classification accuracy, respectively, and quantitatively compare our model with existing advanced methods to show its advantages in multi-source remote sensing data fusion. The experimental results show that the optimized model is superior to the traditional method in multiple evaluation indicators, which verifies its effectiveness and superiority in practical applications.

3.2 Implementation Details

Before training, we randomly crop all training samples, cropping each pair of images into patches of size 128×128 to fit the network input size. This preprocessing method helps to enhance the generalization ability of the model and reduce the computational overhead during training. In the training process of the model, we adopted a training period of 120 epochs and divided the training into three stages in order to gradually optimize the performance of the model. In order to ensure the stability of training, we set the batch size to 4, which means that 4 images are used for training in each iteration, which can effectively control the occupation of video memory and ensure the stability of gradient update. For the optimizer, the AdamW optimizer was chosen, which combines the advantages of Adam and Weight Decay and is better able to handle the overfitting problem during training. The initial learning rate is set to 10-4, and the learning rate decay is performed every 20 iterations with a decay factor of 0.5. This learning rate scheduling strategy helps to achieve finer optimization later in training and avoid overfitting. All experiments are performed under the PyTorch framework, which leverages its powerful deep learning support for model training and debugging.

3.3 Evaluation Metrics

To objectively assess the performance of our proposed infrared and visible image fusion method, we adopt a comprehensive set of widely used evaluation metrics. These metrics are designed to quantify different aspects of fusion quality, including information richness, structural similarity, visual fidelity, and detail preservation. Below is a detailed description of each metric used in our experiments. Entropy measures the amount of information contained in the fused image. Standard deviation reflects the contrast and clarity of the fused image. Spatial frequency evaluates the overall activity level or texture richness in the image. It combines horizontal and vertical frequency components to reflect the sharpness and edge information. SCD measures the correlation difference between the source images and the fused image. Mutual information quantifies the amount of information shared between the fused image and each source image. VIF evaluates how well the fused image preserves the visual information perceived by the human eye. It considers the distortion of natural scene statistics and provides a perceptual measure of fusion quality. These metrics collectively provide a multi-dimensional evaluation of fusion performance, allowing us to compare our method with existing approaches in terms of both objective quality and subjective visual perception.

3.4 Experiment Results

To ensure the reliability and comprehensiveness of our evaluation, we compared our method with several state-of-the-art image fusion models, including DATFuse (Tang et al., 2023), DSFusion (K. Liu et al., 2024), SDCFusion (X. Liu et al., 2024), SeAFusion (Tang et al., 2022), IFCNN (Zhang et al., 2020), UNFusion (Wang et al., 2022), and GAN-FM (Zhang et al., 2021). These models represent advanced approaches in the current image fusion field, encompassing various techniques and frameworks. For example, DATFuse and DSFusion adopt adaptive and multi-scale fusion strategies, respectively, aiming to optimize detail preservation and contrast enhancement; SDCFusion leverages structural content differences to improve fusion performance; SeAFusion and IFCNN focus on applying deep learning models to multisource remote sensing data fusion, further enhancing the quality and information fidelity of fused images; UNFusion employs convolutional neural networks for image fusion, exhibiting strong learning capabilities; GAN-FM utilizes a generative adversarial network (GAN) framework to improve the visual quality and structural fidelity of fused images. By comparing with these models, we comprehensively evaluate the proposed method and further validate its advantages in infrared and visible image fusion.

The fusion results on the TNO dataset are shown in Fig. 2, the quantitative results are shown as Tab1. The fusion results on the RoadScene dataset are shown in Fig. 3, the quantitative results are shown as Tab2, Blodface and underline show the best and second-best values, respectively. The experiments demonstrate that our method achieves excellent visual performance, effectively preserving both rich texture details and salient information. Meanwhile, the quantitative results also show promising outcomes.

Figure 2: Qualitative comparison on 02 from TNO dataset

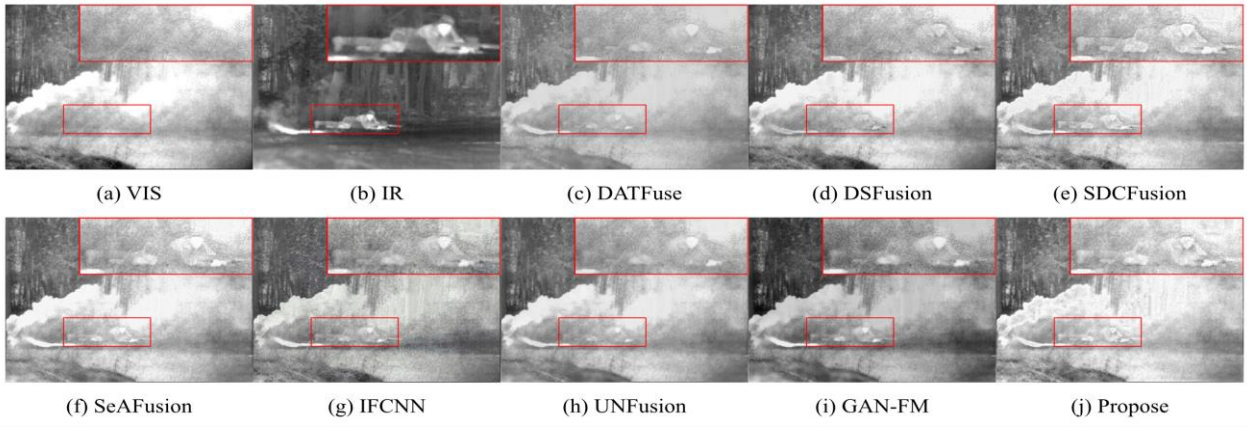


Figure 3: Qualitative comparison on FLIR_09624 from RoadScene dataset

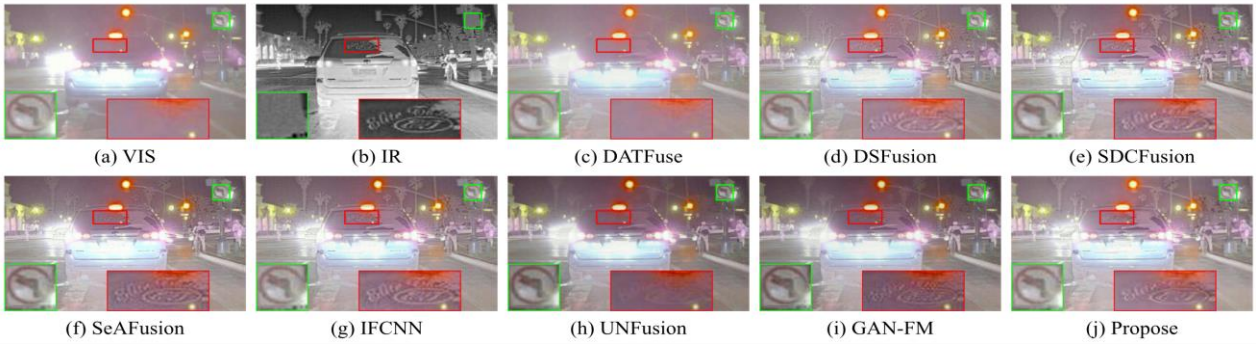


Table 1: Quantitative results on TNO

Methods	EN	SD	SF	SCD	MI	VIF
DATFuse	6.85	27.56	9.58	1.50	3.13	0.62
DSFusion	6.81	34.83	8.99	1.52	2.73	0.65
GAN-FM	6.81	**46.76**	12.50	1.67	3.17	0.58
SeAFusion	**7.13**	43.24	12.25	1.70	2.84	0.71
IFCNN	6.85	37.08	12.95	**1.78**	2.05	0.62
SDCFusion	7.06	39.96	12.09	1.73	2.47	0.64
UNFusion	7.01	41.82	10.05	1.68	**3.81**	0.61
Ours	7.08	43.45	**13.36**	1.65	2.01	**0.72**

Table 2: Quantitative results on 50 random RoadScene data sets

Methods	EN	SD	SF	SCD	MI	VIF
DATFuse	6.73	32.29	11.18	1.29	2.07	0.62
DSFusion	7.01	39.94	13.81	1.31	1.93	0.55
GAN-FM	**7.43**	**52.88**	13.93	1.67	1.99	**0.69**
SeAFusion	7.42	50.56	15.44	1.67	2.05	0.68
IFCNN	7.10	39.21	14.59	1.61	1.95	0.62
SDCFusion	7.28	44.96	**16.21**	1.67	1.90	0.65
UNFusion	7.27	48.33	12.45	**1.70**	2.03	0.61
Ours	7.33	49.84	15.93	1.68	**2.28**	**0.69**

3.5 Experimental Analysis and Summary

From the qualitative results in Fig. 2, it can be clearly observed that our method achieves superior performance in preserving salient objects—especially for human figures. While SDCFusion also performs well in highlighting salient targets, its edge definition is less sharp compared to our approach. On the TNO dataset, our method demonstrates strong performance across multiple evaluation metrics, particularly in Spatial Frequency (SF) and Visual Information Fidelity (VIF) . Although our method does not achieve the highest Entropy (EN) value, it ranks second only to SeAFusion, indicating that our fused images maintain rich information content while preserving fine details effectively. In Fig. 3, our method also shows clear advantages: the infrared-salient text remains highly legible, and the visual information of the signboard is well preserved, outperforming most competing methods visually. The quantitative results on the RoadScene dataset further confirm the superiority of our approach, especially in terms of SF and VIF scores, which reflect the enhanced texture richness and perceptual quality of the fused images. In summary, both qualitative and quantitative evaluations demonstrate that our method excels in preserving image details, enhancing structural clarity, and maintaining visual fidelity, making it a promising solution for infrared and visible image fusion tasks.

4. Conclusion

This paper proposes a local feature extractor SDI based on multi-window attention and residual connections, alongside a global feature extractor GFE that integrates convolutional operations with Transformer architecture. This design enables the fused image to effectively preserve salient target

information while simultaneously capturing rich texture details. Experimental results on multiple benchmark datasets demonstrate that our fusion method achieves superior performance in terms of image quality, detail preservation, and visual consistency compared to existing techniques. However, despite the outstanding performance across several datasets, there remain issues worthy of further investigation. Specifically, the relatively complex model architecture results in higher computational costs during training and inference, which may limit its applicability in scenarios requiring real-time processing. In future work, we plan to incorporate knowledge distillation techniques to enhance efficiency and performance.

References

Liu, K., Li, M., Chen, C., Rao, C., Zuo, E., Wang, Y., Yan, Z., Wang, B., Chen, C., & Lv, X. (2024). DSFusion: Infrared and visible image fusion method combining detail and scene information. *Pattern Recognition, 154*, Article 110633. https://doi.org/10.1016/J.PATCOG.2024.110633

Liu, X., Huo, H., Li, J., Pang, S., & Zheng, B. (2024). A semantic-driven coupled network for infrared and visible image fusion. *Information Fusion, 108*, Article 102352. https://doi.org/10.1016/J.INFFUS.2024.102352

Liu, Y., Liu, S., & Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Information Fusion, 24*, 147-164. https://doi.org/10.1016/J.INFFUS.2014.09.004

Ma, J., Ma, Y., & Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. *Information Fusion, 45*, 153-178. https://doi.org/10.1016/J.INFFUS.2018.02.004

Ma, J., Yu, W., Liang, P., Li, C., & Jiang, J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion, 48*, 11-26. https://doi.org/10.1016/J.INFFUS.2018.09.004

Park, S. C., Park, M. K., & Kang, M. G. (2003). Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine, 20*(3), 21-36. https://doi.org/10.1109/MSP.2003.1203207

Pizurica, A., Philips, W., Lemahieu, I., & Acheroy, M. (2003). A versatile wavelet domain noise filtration technique for medical imaging. *IEEE Transactions on Medical Imaging, 22*(3), 323-331. https://doi.org/10.1109/TMI.2003.809588

Tang, L., Yuan, J., & Ma, J. (2022). Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion, 82*, 28-42. https://doi.org/10.1016/J.INFFUS.2021.12.004

Tang, W., He, F., Liu, Y., Duan, Y., & Si, T. (2023). DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(7), 3159-3172. https://doi.org/10.1109/TCSVT.2023.3234340

Toet, A. (1989). Image fusion by a ration of low-pass pyramid. *Pattern Recognition Letters, 9*(4), 245-253. https://doi.org/10.1016/0167-8655(89)90003-2

Toet, A. (2017). The TNO multiband image data collection. *Data in Brief, 15*, 249-251. https://doi.org/10.1016/J.DIB.2017.09.038

Wang, Z., Wang, J., Wu, Y., Xu, J., & Zhang, X. (2022). UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(6), 3360-3374. https://doi.org/10.1109/TCSVT.2021.3109895

Xu, H., Ma, J., Jiang, J., Guo, X., & Ling, H. (2022). U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 502-518. https://doi.org/10.1109/TPAMI.2020.3012548

Zhang, H., Yuan, J., Tian, X., & Ma, J. (2021). GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual markovian discriminators. *IEEE Transactions on Computational Imaging, 7,* 1134-1147. https://doi.org/10.1109/TCI.2021.3119954

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., & Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion, 54*, 99-118. https://doi.org/10.1016/J.INFFUS.2019.07.011

Zhao, W., Xie, S., Zhao, F., He, Y., & Lu, H. (2023). *MetaFusion: Infrared and visible image fusion via meta-feature embedding from object detection* [Paper presentation]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada.

Zhou, Y., Fu, J., Chen, Z., Zhuge, F., Wang, Y., Yan, J., Ma, S., Xu, L., Yuan, H., Chan, M., Miao, X., He, Y., & Chai, Y. (2023). Computational event-driven vision sensors for in-sensor spiking neural networks. *Nature Electronics, 6*(11), 870-878. https://doi.org/10.1038/S41928-023-01055-2

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights