Wavelet-Assisted Multi-Frequency Attention Network for Pansharpening

Jie Huang[1*], **Rui Huang**[1*], **Jinghao Xu**[1], **Siran Peng**[2, 3], **Yule Duan**[1], **Liangjian Deng**[1†]

[1]University of Electronic Science and Technology of China
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
{jayhuang,huang_rui,2023040901014}@std.uestc.edu.cn,pengsiran2023@ia.ac.cn,
yule.duan@outlook.com,liangjian.deng@uestc.edu.cn

Abstract

Pansharpening aims to combine a high-resolution panchromatic (PAN) image with a low-resolution multispectral (LRMS) image to produce a high-resolution multispectral (HRMS) image. Although pansharpening in the frequency domain offers clear advantages, most existing methods either continue to operate solely in the spatial domain or fail to fully exploit the benefits of the frequency domain. To address this issue, we innovatively propose Multi-Frequency Fusion Attention (MFFA), which leverages wavelet transforms to cleanly separate frequencies and enable lossless reconstruction across different frequency domains. Then, we generate Frequency-Query, Spatial-Key, and Fusion-Value based on the physical meanings represented by different features, which enables a more effective capture of specific information in the frequency domain. Additionally, we focus on the preservation of frequency features across different operations. On a broader level, our network employs a wavelet pyramid to progressively fuse information across multiple scales. Compared to previous frequency domain approaches, our network better prevents confusion and loss of different frequency features during the fusion process. Quantitative and qualitative experiments on multiple datasets demonstrate that our method outperforms existing approaches and shows significant generalization capabilities for real-world scenarios.

Code — https://github.com/Jie-1203/WFANet

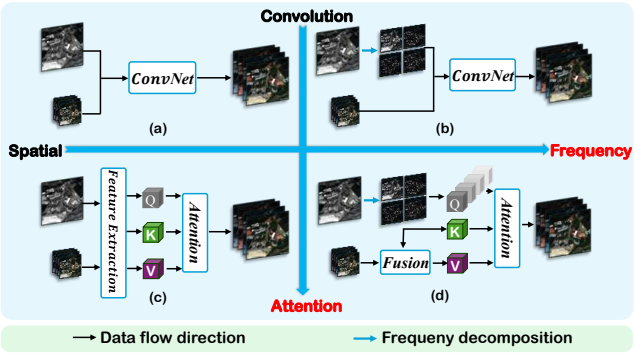

Figure 1: The comparison covers four methods across two dimensions: (a) Convolutional network in the spatial domain, (b) Convolutional network in different frequency domains, (c) Attention mechanism in the spatial domain, and (d) Our proposed method which forms the primary motivation for this paper: 1) utilizing wavelet transforms to process in different frequency domains; 2) designing an attention method with clear physical significance to leverage the advantages of frequency domain processing.

Introduction

High-resolution multispectral (HRMS) images are vital for applications like environmental monitoring and urban planning. Due to hardware constraints, satellites typically capture low-resolution multispectral (LRMS) and high-resolution panchromatic (PAN) images. Pansharpening fuses these to produce HRMS, combining their strengths to enhance spatial and spectral resolution.

To obtain high-resolution multispectral (HRMS) images, pansharpening methods are categorized into traditional and deep learning-based approaches. Traditional methods are divided into three groups (Meng et al. 2019): Component Substitution (CS) (Vivone 2019), Multi-Resolution Analysis (MRA) (Vivone, Restaino, and Chanussot 2018), and Variational Optimization-based (VO) (Tian et al. 2022) techniques. In recent years, with the rapid development of deep learning, many deep learning methods (Wang et al. 2021; Zhang et al. 2023) have been proposed for pansharpening using convolutional neural networks (CNN), as shown in Fig. 1 (a), such as PNN (Masi et al. 2016), DiCNN (He et al. 2019), and LAGConv (Jin et al. 2022b). These methods underscore deep learning's potential to improve pansharpening accuracy and efficiency. However, most existing methods do not process images in different frequency domains but instead operate in the original single spatial domain, thereby limiting the potential for improving fusion quality.

Direct fusion in the spatial domain methods can often result in detail loss or blurring due to the imprecise separation of frequency information. In contrast, frequency-based methods can separate different frequencies for targeted processing, which better preserves hard-to-capture high-frequency information while effectively preventing interference between different frequencies. Consequently, pro-

[*]Co-first authors contributed equally.
[†]Corresponding author.

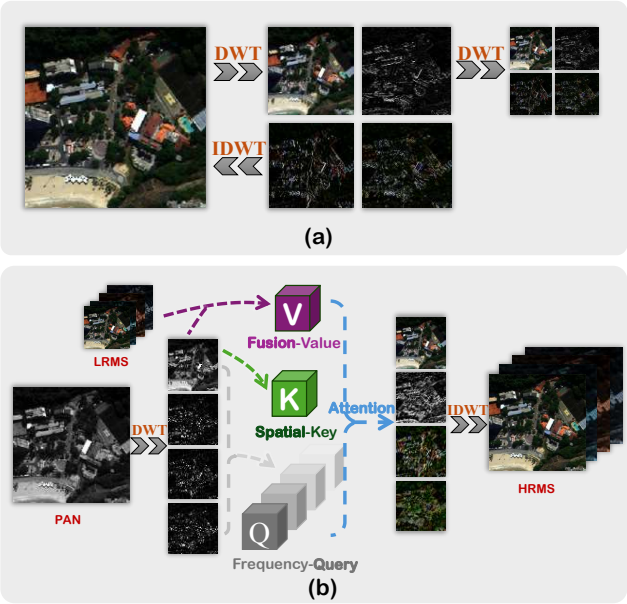

Figure 2: (a) DWT decomposes the image into four different frequency components. IDWT is the lossless inverse process of DWT. Multiple applications of DWT produce a multi-scale wavelet pyramid. (b) Simplified illustration of MFFA. Fusion-Value, Spatial-Key, and Frequency-Query are derived from the information indicated by the arrows. These components are then processed through an attention mechanism, enabling the reconstruction of features across different frequencies that integrate both spectral and spatial information.

cessing in different frequency domains can be a more effective approach for achieving better fusion results compared to spatial domain fusion methods, and some works (Jin et al. 2022a; Ran et al. 2023) have already attempted this approach. For example, AFM-DIN (Li et al. 2022) introduces a high-frequency injection module to enhance LRMS features with PAN details, but it may lose low-frequency information. FAMENet (He et al. 2024b) uses an expert mixture model to fuse different frequencies, effectively balancing both high-frequency and low-frequency information. However, these methods often struggle to achieve clean separation, leading to interference and information loss due to the neural network's tendency to slightly blend frequency components together (Shan, Li, and Wang 2021). To address these issues, we adopt a method that cleanly separates frequencies and enables lossless reconstruction. As shown in Fig. 2 (a), the Discrete Wavelet Transform (DWT) cleanly separates frequency components (Mallat 1989; Fujieda, Takayama, and Hachisuka 2018). The Inverse Discrete Wavelet Transform (IDWT) is lossless, preserving all information. Repeated DWT builds a wavelet pyramid (Liu et al. 2018), efficiently handling multi-scale features and enhancing detail detection, offering advantages over other methods that attempt to separate frequency components.

Using wavelet transforms to fuse information across different frequency domains is an innovative approach. Designing appropriate networks and modules to effectively extract and combine these features is therefore crucial for achieving optimal fusion results. Some past methods have employed wavelet transforms for pansharpening, as shown in Fig. 1 (b), such as FAFNet (Xing et al. 2023), which utilizes DWT layers to extract and manage frequency-domain features, followed by IDWT layers that reconstruct these features into the spatial domain, with the final fusion achieved through a convolutional block that produces the high-quality fused image. However, convolutional neural networks inherently excel at capturing low-frequency information and are less effective in capturing high-frequency details (Xu et al. 2019; Yedla and Dubey 2021). This limitation leads to decreased fusion quality. To address this issue, we consider leveraging attention mechanisms that can more flexibly capture specific features (Soydaner 2022). Some previous methods (Hou et al. 2024; Deng et al. 2023) have attempted to use attention mechanisms in the spatial domain, as shown in Fig. 1 (c). For example, PanFormer (Zhou, Liu, and Wang 2022) employs a customized Transformer architecture, using panchromatic (PAN) and low-resolution multispectral (LRMS) features as queries and keys for joint feature learning, modeling long-range dependencies to produce high-quality pan-sharpened results. However, such a design does not explicitly capture and fuse information from different frequencies. In other visual domains, combining wavelet transforms with attention mechanisms, such as Wave-ViT (Yao et al. 2022), integrates wavelet transforms with Transformers. By performing invertible down-sampling within Transformer blocks, this method improves image recognition and enhances visual representation accuracy.

Inspired by the above discussion, we design an innovative Wavelet-Assisted Multi-Frequency Attention Network called WFANet, *with the core component being the Multi-Frequency Fusion Attention (MFFA).* In our proposed MFFA, *we introduce the concept of a Frequency Attention Triplet, which consists of Frequency-Query, Spatial-Key, and Fusion-Value*. As shown in Fig. 2 (b), we apply DWT to the PAN image, achieving a clean separation of different frequency features, effectively avoiding interference between them. To process and fuse this separated frequency information, unlike conventional attention mechanisms (Vaswani et al. 2017; Soydaner 2022), our Frequency Attention Triplet has distinct physical significance: Frequency-Query represents frequency features, Spatial-Key encodes spatial information, and Fusion-Value represents the preliminary fusion of spatial and spectral features. This design effectively guides the information fusion process. The attention mechanism then captures correlations to achieve the initial fusion of frequency features, and IDWT is used for lossless reconstruction. Through MFFA, we can more effectively fuse information across different frequency domains and prevent the loss of frequency. Additionally, inspired by previous methods (Deng et al. 2021; Hou et al. 2023; Wu et al. 2020), which demonstrated that enhancing spatial details significantly improves restoration when a module is primarily focused on fusion, as realized in our MFFA, *we design the Spatial Detail Enhancement Module (SDEM) to focus on the*

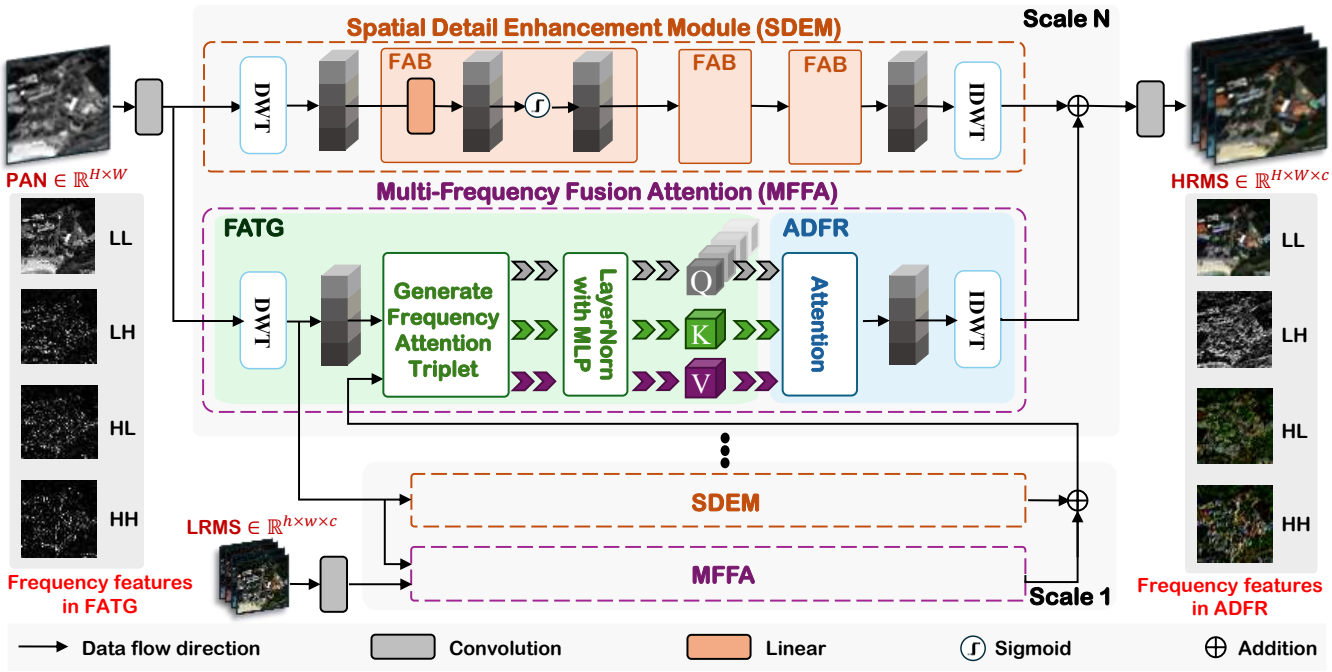

Figure 3: The overall workflow of our WFANet. Our network processes the data using multiple scales (only two scales are illustrated here for simplicity). WFANet consists of two sub-modules: the Multi-Frequency Fusion Attention (MFFA) and the Spatial Detail Enhancement Module (SDEM). The illustration of frequency features is shown on both sides of the figure.

extraction and enhancement of spatial details. In designing SDEM, we compare different operations for their adaptability to the frequency domain, ensuring better preservation of spatial details. Moreover, our overall framework is a multi-scale progressive reconstruction framework, fully utilizing the inherent multi-scale nature of the wavelet pyramid.

In summary, the contributions of this work are as follows:

- We introduce the Multi-Frequency Fusion Attention (MFFA), utilizing wavelet transforms to cleanly separate and accurately reconstruct frequency components. This approach integrates the Frequency-Query, Spatial-Key, and Fusion-Value triplet to enhance feature fusion precision across different frequency domains, effectively reducing confusion and information loss.
- Additionally, we focus on how different operations preserve frequency features and utilize the wavelet pyramid for progressive, multi-scale fusion. The effectiveness of these strategies has been validated and demonstrated through extensive ablation experiments.
- Our method achieves state-of-the-art performance on three diverse pansharpening datasets, demonstrating high-quality fusion results supported by both quantitative and qualitative experimental evidence.

Proposed Method

This section introduces the proposed WFANet, detailing its two key components: Multi-Frequency Fusion Attention (MFFA) and the Spatial Detail Enhancement Module (SDEM), followed by the overall multi-scale framework. Fig. 3 illustrates the workflow of WFANet.

Multi-Frequency Fusion Attention (MFFA)

To fuse information across frequencies, we propose the MFFA, which is composed of two phases: Frequency Attention Triplet Generation (FATG) and Attention-Driven Frequency Reconstruction (ADFR). Details are shown in Fig. 4.

(I) Frequency Attention Triplet Generation As in the typical attention mechanism (Vaswani et al. 2017; Soydaner 2022) , Query, Key, and Value are the three components of attention. Query represents the information we seek, Key is the index of this information, and Value is the specific content. To better adapt to different frequency domains, we design the Frequency Attention Triplet. Specifically, different frequency features are the information we seek, the overall spatial features are the indices for querying different frequency features, and the specific content is the fusion of spectral and spatial information. Therefore, we design Frequency-Query, Spatial-Key, and Fusion-Value with specific physical meanings. We first perform a DWT on P, which is the feature of the panchromatic (PAN) image after convolution, as shown below:

$$P_{LL}, P_{LH}, P_{HL}, P_{HH} = \mathrm{DWT}(P) \tag{1}$$

where P_{LL} represents the low-frequency details of P, and P_{LH}, P_{HL}, and P_{HH} represent the high-frequency details of P in the horizontal, vertical, and diagonal directions, respectively. *For convenience,* $i = LL, LH, HL, HH$ *cor-*

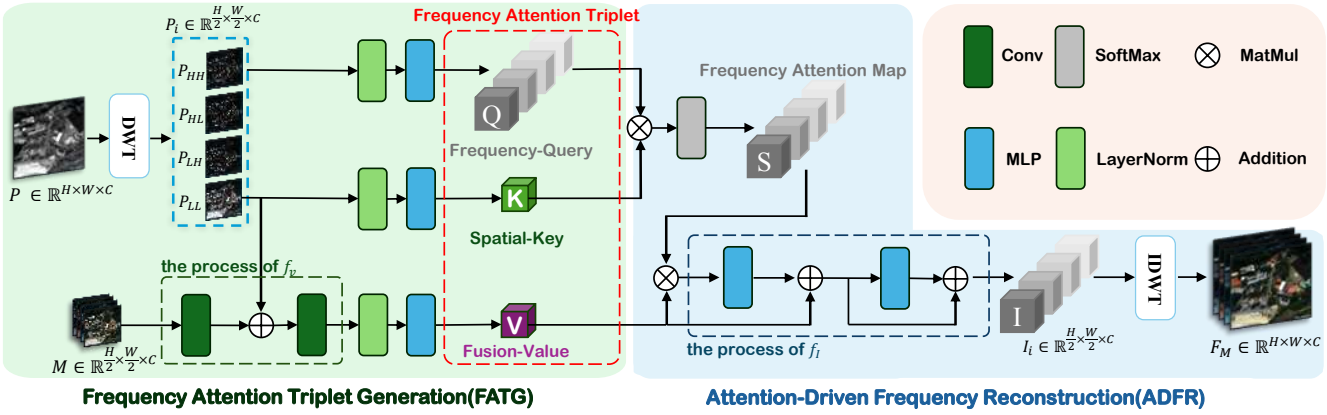

Figure 4: The MFFA workflow involves two phases. First, in the FATG phase, the Frequency Attention Triplet with specific physical significance is generated. Then, in the ADFR phase, the obtained Frequency Attention Triplet is processed to reconstruct the features at different frequencies. Q, S, and I are shown as four colored blocks, representing features from different frequency domains after DWT. The data dimensions are exemplified using the largest scale.

responds to the four frequency features mentioned above, respectively. The DWT operation has already separated the frequency features adequately, thus we directly use P_i, representing different frequency features, as Q_i. Low-frequency features represent the overall spatial appearance of the image, while high-frequency information represents edge details and fine textures (Kingsbury and Magarey 1998). Therefore, to use the overall spatial features as our key, we directly take the low-frequency features represented by P_{LL} as the Spatial-Key. Then, we design an f_v to fuse spatial and spectral information as the Fusion-Value. Next, the three components are normalized using LayerNorm and processed through an MLP to enhance their expressiveness. The specific process can be described by the following equations:

$$\begin{aligned} Q_i &= \mathrm{MLP}(\mathrm{LN}(P_i)) \\ K &= \mathrm{MLP}(\mathrm{LN}(P_{LL})) \\ V &= \mathrm{MLP}(\mathrm{LN}(f_v(M, P_{LL}))) \end{aligned} \tag{2}$$

where M represents the feature of the low-resolution multispectral (LRMS) image after convolution and f_v represents the process of obtaining the Fusion-Value by combining M and P_{LL} through convolution.

(II) Attention-Driven Frequency Reconstruction After obtaining the Frequency Attention Triplet, we use the attention mechanism to reconstruct the fused features at different frequencies. First, calculate the frequency correlation R_i between the Frequency-Query and Spatial-Key, representing the correlation between different frequency features and the overall spatial features. Then, apply softmax to obtain the Frequency Attention Map S_i, which highlights the importance of different frequency features relative to the overall spatial features. The process is as follows:

$$\begin{aligned} R_i &= Q_i \otimes K \\ S_i &= \mathrm{softmax}(R_i) \end{aligned} \tag{3}$$

where $\otimes$ represents matrix multiplication and softmax is an operation that converts input values into a probability distribution. Next, the Frequency Attention Map S_i of different frequencies is separately multiplied by V, which is the fusion of spectral information and low-frequency spatial information. Then, the result is processed through MLPs and residual connection to obtain the reconstructed features of different frequencies containing spectral information I_i. This process can be described by the following equation:

$$I_i = f_I(S_i \otimes V) \tag{4}$$

where f_I represents the process of applying MLPs and residual connection to the result of the multiplication. Finally, the preliminary reconstructed image F_M is obtained by leveraging the lossless property of the IDWT, as shown below:

$$F_M = \mathrm{IDWT}(I_{LL}, I_{LH}, I_{HL}, I_{HH}) \tag{5}$$

where I_{LL}, I_{LH}, I_{HL}, and I_{HH} represent the features reconstructed at different frequencies.

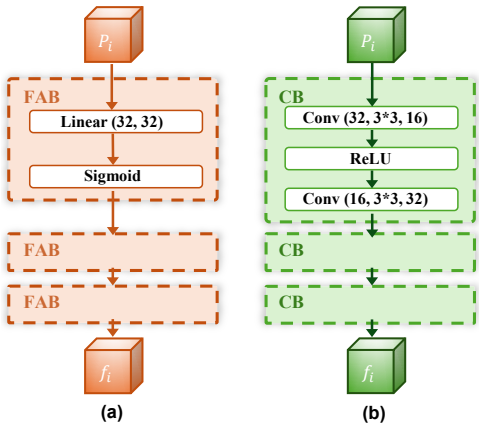

Figure 5: Comparison of two network architectures for the SDEM: (a) Frequency Adaptation Block (FAB), which is used in the SDEM. (b) Convolution Block (CB).

Spatial Detail Enhancement Module (SDEM)

The core component, MFFA, achieves the fusion of information across different frequency domains. In contrast, SDEM focuses on extracting and enhancing spatial detail information within these frequency domains. First, we decompose P according to Eq.1 to obtain P_i, features containing information from different frequencies, helping to prevent interference between them. Next, we extract f_i, representing spatial information for different frequency features, separately using several Frequency Adaptation Blocks (FABs). An FAB is a block capable of adapting to different frequencies and is composed of a linear layer and a sigmoid activation function. This process is illustrated below:

$$f_i = \mathrm{FABs}(P_i) \tag{6}$$

where $i = LL, LH, HL, HH$ correspond to the four different frequency features, respectively. As illustrated in Fig. 5, we do not choose the Convolution Block. Given that convolutional networks struggle with extracting high-frequency information and perform poorly when processing different frequency domains (Xu et al. 2019; Yedla and Dubey 2021), we opt for linear layers, which, as demonstrated by our ablation experiments, better adapt to different frequency domains. After extracting features in different frequency domains, we use the lossless IDWT to recover the complete spatial details F_S, as illustrated below:

$$F_S = \mathrm{IDWT}(f_{LL}, f_{LH}, f_{HL}, f_{HH}) \tag{7}$$

Network Framework and Loss

This section describes how to utilize the wavelet pyramid to construct the multi-scale network architecture of WFANet with MFFA and SDEM. Our network employs a multi-scale structure with N layers (limited by the dataset, we use two scales in this paper). First, we construct a wavelet pyramid by repeatedly applying DWT, as follows:

$$P_k = \mathrm{DWT}(P_{k+1}) \tag{8}$$

where P_{k+1} represents the four different frequency features of the previous larger scale, and P_k represents the frequency features of the next smaller scale. Then, we progressively fuse from the smallest scale. P_k and M_k are the inputs at the k-th scale. The output of each layer is obtained by adding the output F_M from the MFFA and the output F_S from the SDEM. The output of the k-th layer serves as the input M_{k+1} for the next layer, which in turn enables the process of progressive reconstruction, expressed as follows:

$$\begin{cases} M_1 = f(M_0, P_0) \\ M_2 = f(M_1, P_1) \\ \quad\vdots \\ M_n = f(M_{n-1}, P_{n-1}) \end{cases} \tag{9}$$

where n represents the number of scales. M_n is then convolved to get the final high-resolution multispectral (HRMS) image $\hat{M}$. We choose the simple ℓ_1 loss function since it is sufficient to yield consistently good outcomes:

$$\mathcal{L} = \frac{1}{K}\sum_{i=1}^{K} \|\hat{M}^{\{i\}} - I^{\{i\}}\|_1 \tag{10}$$

where K is the number of training data, $I^{\{i\}}$ denotes the i-th ground truth image, and $\|\cdot\|_1$ represents the ℓ_1 norm.

Experiments

Datasets and Benchmark

To benchmark the effectiveness of our network for pansharpening, we adopt various datasets, including datasets captured by the WorldView-3 (WV3), GaoFen-2 (GF2), and QuickBird (QB) sensors. Since ground truth (GT) images are not available, Wald's protocol (Wald, Ranchin, and Mangolini 1997) is applied. Each training dataset consists of PAN, LRMS, and GT image pairs with sizes of 64 × 64, 16 × 16 × 8, and 64 × 64 × 8, respectively. We obtain our datasets and data processing methods from the PanCollection repository[1] (Deng et al. 2022). To evaluate the proposed method, several state-of-the-art pansharpening methods are selected, including three traditional methods, MTF-GLP-FS (Vivone, Restaino, and Chanussot 2018), BDSD-PC (Vivone 2019), and TV (Palsson, Sveinsson, and Ulfarsson 2013), and seven deep-learning methods, including PNN, PanNet (Yang et al. 2017), DiCNN, FusionNet (Deng et al. 2021), U2Net (Peng et al. 2023), PanMamba (He et al. 2024a), and CANNet (Duan et al. 2024).

Implementation Details

We implement our network using the PyTorch framework on an RTX 4090D GPU. The learning rate is set to 9×10^{-4} and is halved every 90 epochs. The model is trained for 360 epochs with a batch size of 32. The Adam optimizer is employed. Our method's performance is assessed using standard pansharpening metrics including SAM (Boardman 1993), ERGAS (Wald 2002), and Q4/Q8 (Garzelli and Nencini 2009) for reduced-resolution datasets, and HQNR (Arienzo et al. 2022), D_s, and D_λ for full-resolution datasets.

Comparison with State-of-the-Art Methods

Reduced-Resolution Assessment Table 1 clearly shows a comparison of our proposed method with the current best methods across three datasets. Our method consistently achieves the best results across all metrics. Specifically, our method achieves a PSNR improvement of 0.228dB, 0.334dB, and 0.417dB on the WV3, QB, and GF2 datasets, respectively, compared to the second-best results. These improvements highlight the clear advantages of our method, confirming its competitiveness in the field. Fig. 6 and Fig. 7 show the qualitative assessment results for two datasets and their corresponding ground truth (GT). By comparing the MSE residuals between the pan-sharpened results and the ground truth, it is evident that our residual maps are the darkest, indicating that our method outperforms others. The experimental results above demonstrate that our method is superior to the latest state-of-the-art pansharpening methods.

[1] https://github.com/liangjiandeng/PanCollection

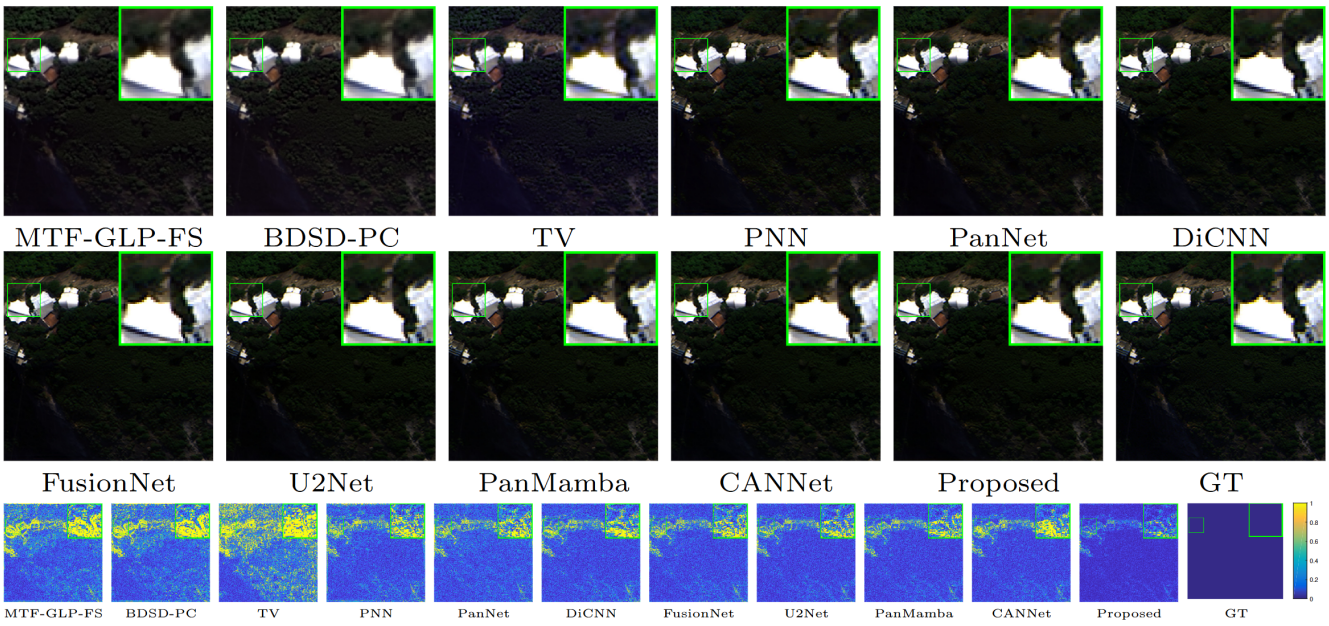

Figure 6: The visual results (Top) and residuals (Bottom) of all compared approaches on the WV3 reduced-resolution dataset.

Methods	WV3				QB				GF2			
	PSNR↑	SAM↓	ERGAS↓	Q8↑	PSNR↑	SAM↓	ERGAS↓	Q4↑	PSNR↑	SAM↓	ERGAS↓	Q4↑
MTF-GLP-FS	32.963	5.316	4.700	0.833	32.709	7.792	7.373	0.835	35.540	1.655	1.589	0.897
BDSD-PC	32.970	5.428	4.697	0.829	32.550	8.085	7.513	0.831	35.180	1.681	1.667	0.892
TV	32.381	5.692	4.855	0.795	32.136	7.510	7.690	0.821	35.237	1.911	1.737	0.907
PNN	37.313	3.677	2.681	0.893	36.942	5.181	4.468	0.918	39.071	1.048	1.057	0.960
PanNet	37.346	3.613	2.664	0.891	34.678	5.767	5.859	0.885	40.243	0.997	0.919	0.967
DiCNN	37.390	3.592	2.672	0.900	35.781	5.367	5.133	0.904	38.906	1.053	1.081	0.959
FusionNet	38.047	3.324	2.465	0.904	37.540	4.904	4.156	0.925	39.639	0.974	0.988	0.964
U2Net	39.117	2.888	2.149	0.920	38.065	4.642	3.987	0.931	43.379	0.714	0.632	0.981
PanMamba	39.012	2.913	2.184	0.920	37.356	4.625	4.277	0.929	42.907	0.743	0.684	0.982
CANNet	39.003	2.941	2.174	0.920	38.488	4.496	3.698	0.937	43.496	0.707	0.630	0.983
Proposed	**39.345**	**2.849**	**2.093**	**0.922**	**38.822**	**4.392**	**3.556**	**0.940**	**43.913**	**0.685**	**0.597**	**0.985**

Table 1: Comparisons on WV3, QB, and GF2 datasets with 20 reduced-resolution samples, respectively. Best: **bold**, and second-best: underline.

Ablation	PSNR↑	SAM↓	ERGAS↓	Q8↑
Frequency-Query	38.884	2.968	2.206	0.918
Spatial-Key	39.156	2.901	2.138	0.921
Fusion-Value	38.921	2.971	2.203	0.918
Ours	**39.345**	**2.849**	**2.093**	**0.922**

Table 3: Ablation experiment about Attention Triplet on WV3 reduced-resolution dataset.

Full-Resolution Assessment To demonstrate the generalization ability of our method, we conduct experiments on full-resolution samples of WV3. The quantitative evaluation results are shown in Table 2. Our method achieves the best HQNR results, which reflects its ability to balance spectral and spatial distortions, demonstrating its high application value.

Ablation Study

This section explores the rationale behind the design of the Frequency Attention Triplet and the roles of key components and strategies in WFANet. We conduct a series of ablation experiments on the WV3 dataset to demonstrate their effectiveness and validity. First, Table 3 presents three sets of ablation experiments for Frequency Attention Triplet, followed by Table 4, which shows four sets of ablation experiments for WFANet. *More experiments and detailed ablation settings can be found in the supplementary materials.*

Frequency Attention Triplet To demonstrate the effectiveness of Frequency-Query, we remove the DWT operation and directly use a spatial domain Q instead of using Q_i in the different frequency domains. For Spatial-Key, we no longer use P_{LL} obtained by DWT, which represents the overall spatial features, but instead use features from P after

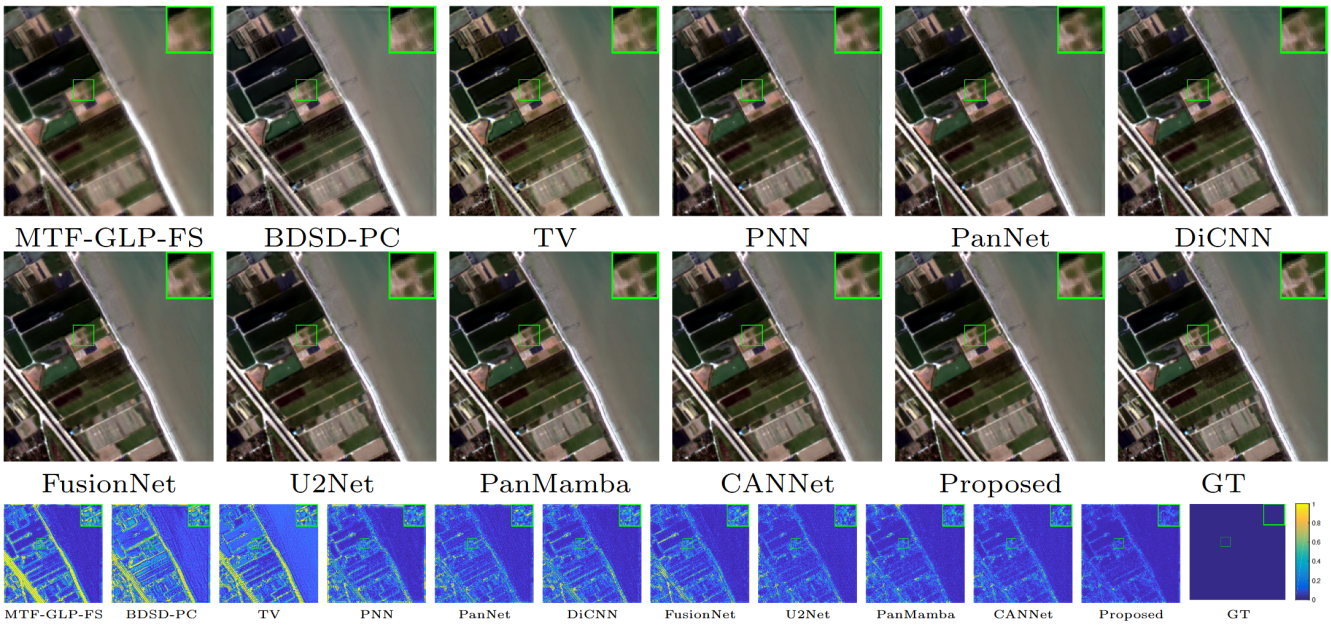

Figure 7: The visual results (Top) and residuals (Bottom) of all compared approaches on the GF2 reduced-resolution dataset.

Metric	MTF-GLP-FS	BDSD-PC	TV	PNN	PanNet	DiCNN	FusionNet	U2Net	PanMamba	CANNet	Proposed
$D_\lambda \downarrow$	0.020	0.063	0.023	0.021	**0.017**	0.036	0.024	0.020	0.018	0.020	**0.017**
$D_s \downarrow$	0.063	0.073	0.039	0.043	0.047	0.046	0.036	0.028	0.053	0.030	**0.027**
HQNR↑	0.919	0.870	0.938	0.937	0.937	0.920	0.941	0.952	0.930	0.951	**0.957**

Table 2: Quantitative comparisons on the WV3 full-resolution dataset.

Ablation	PSNR↑	SAM↓	ERGAS↓	Q8↑
MFFA	38.486	3.148	2.349	0.915
SDEM	38.993	2.937	2.178	0.918
Multi-Scale	38.851	2.995	2.214	0.916
FAB	39.074	2.919	2.155	0.919
Ours	**39.345**	**2.849**	**2.093**	**0.922**

Table 4: Ablation experiment about key components and strategy on WV3 reduced-resolution dataset.

convolution, introducing interference from high-frequency information. Regarding Fusion-Value, we no longer use the fusion of LRMS and P_{LL}, but instead include only the information from LRMS, thereby lacking the low-frequency spatial information represented by P_{LL}. The results in Table 3 demonstrate the effectiveness of each component in the Frequency Attention Triplet.

Multi-Frequency Fusion Attention To demonstrate the effectiveness of the attention mechanism in MFFA, we replace it with the convolutional network, where HRMS and different frequency features are concatenated and then processed through convolution. The results in Table 4 prove it.

Spatial Detail Enhancement Module The SDEM enhances reconstructed images by injecting spatial details from frequency domains. We remove the SDEM while retaining the MFFA, and the lack of spatial detail is noticeable. The results in Table 4 confirm the effectiveness of the SDEM.

The Multi-Scale Training Strategy To demonstrate the effectiveness of the multi-scale strategy, we replace the multi-scale network in this paper with a single-scale network, aligning the sizes using convolution operations. The results in Table 4 confirm the significance of this strategy.

Frequency Adaptation Block We explore how operations adapt to frequency domains. Fig. 5 shows two network architectures for the SDEM. We replace the FABs with Convolution Blocks, and Table 4 shows the superior feature extraction across frequencies provided by the FABs.

Conclusion

In this paper, we propose a novel approach, Multi-Frequency Fusion Attention (MFFA), which leverages an effective method for frequency decomposition and reconstruction. By designing Frequency-Query, Spatial-Key, and Fusion-Value with clear physical significance, MFFA achieves more effective and precise fusion in the frequency domain. We also emphasize the adaptation of different operations to the frequency domain and have designed a comprehensive multi-scale fusion strategy. Ablation experiments further confirm the effectiveness of our approach. Extensive experiments on three different satellite datasets demonstrate that our model outperforms state-of-the-art methods.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (12271083), and the Natural Science Foundation of Sichuan Province (2024NSFSC0038).

References

Arienzo, A.; Vivone, G.; Garzelli, A.; Alparone, L.; and Chanussot, J. 2022. Full-Resolution Quality Assessment of Pansharpening: Theoretical and hands-on approaches. *IEEE Geoscience and Remote Sensing Magazine*, 10(3): 168–201.

Boardman, J. W. 1993. Automating spectral unmixing of AVIRIS data using convex geometry concepts. In *JPL, Summaries of the 4th Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*.

Deng, L.-J.; Vivone, G.; Jin, C.; and Chanussot, J. 2021. Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 6995–7010.

Deng, L.-J.; Vivone, G.; Paoletti, M. E.; Scarpa, G.; He, J.; Zhang, Y.; Chanussot, J.; and Plaza, A. 2022. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3): 279–315.

Deng, S.-Q.; Deng, L.-J.; Wu, X.; Ran, R.; and Wen, R. 2023. Bidirectional Dilation Transformer for Multispectral and Hyperspectral Image Fusion. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Duan, Y.; Wu, X.; Deng, H.; and Deng, L.-J. 2024. Content-Adaptive Non-Local Convolution for Remote Sensing Pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27738–27747.

Fujieda, S.; Takayama, K.; and Hachisuka, T. 2018. Wavelet convolutional neural networks. *arXiv preprint arXiv:1805.08620*.

Garzelli, A.; and Nencini, F. 2009. Hypercomplex Quality Assessment of Multi/Hyperspectral Images. *IEEE Geoscience and Remote Sensing Letters*, 662–665.

He, L.; Rao, Y.; Li, J.; Chanussot, J.; Plaza, A.; Zhu, J.; and Li, B. 2019. Pansharpening via Detail Injection Based Convolutional Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1188–1204.

He, X.; Cao, K.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; and Zhou, M. 2024a. Pan-mamba: Effective pan-sharpening with state space model. *arXiv preprint arXiv:2402.12192*.

He, X.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; and Zhou, M. 2024b. Frequency-Adaptive Pan-Sharpening with Mixture of Experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2121–2129.

Hou, J.; Cao, Q.; Ran, R.; Liu, C.; Li, J.; and Deng, L.-j. 2023. Bidomain modeling paradigm for pansharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, 347–357.

Hou, J.; Cao, Z.; Zheng, N.; Li, X.; Chen, X.; Liu, X.; Cong, X.; Zhou, M.; and Hong, D. 2024. Linearly-evolved Transformer for Pan-sharpening. *arXiv preprint arXiv:2404.12804*.

Jin, C.; Deng, L.-J.; Huang, T.-Z.; and Vivone, G. 2022a. Laplacian pyramid networks: A new approach for multispectral pansharpening. *Information Fusion*, 78: 158–170.

Jin, Z.-R.; Zhang, T.-J.; Jiang, T.-X.; Vivone, G.; and Deng, L.-J. 2022b. LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1113–1121.

Kingsbury, N.; and Magarey, J. 1998. Wavelet transforms in image processing.

Li, Y.; Zheng, Y.; Li, J.; Song, R.; and Chanussot, J. 2022. Hyperspectral pansharpening with adaptive feature modulation-based detail injection network. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level Wavelet-CNN for Image Restoration. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Mallat, S. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 674–693.

Masi, G.; Cozzolino, D.; Verdoliva, L.; and Scarpa, G. 2016. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7): 594.

Meng, X.; Shen, H.; Li, H.; Zhang, L.; and Fu, R. 2019. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion*, 102–113.

Palsson, F.; Sveinsson, J. R.; and Ulfarsson, M. O. 2013. A New Pansharpening Algorithm Based on Total Variation. *IEEE Geoscience and Remote Sensing Letters*, 318–322.

Peng, S.; Guo, C.; Wu, X.; and Deng, L.-J. 2023. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 3219–3227.

Ran, R.; Deng, L.-J.; Jiang, T.-X.; Hu, J.-F.; Chanussot, J.; and Vivone, G. 2023. GuidedNet: A General CNN Fusion Framework via High-Resolution Guidance for Hyperspectral Image Super-Resolution. *IEEE Transactions on Cybernetics*, 1–14.

Shan, L.; Li, X.; and Wang, W. 2021. Decouple the high-frequency and low-frequency information of images for semantic segmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1805–1809. IEEE.

Soydaner, D. 2022. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*, 34(16): 13371–13385.

Tian, X.; Chen, Y.; Yang, C.; and Ma, J. 2022. Variational Pansharpening by Exploiting Cartoon-Texture Similarities. *IEEE Transactions on Geoscience and Remote Sensing*, 1–16.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vivone, G. 2019. Robust Band-Dependent Spatial-Detail Approaches for Panchromatic Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 6421–6433.

Vivone, G.; Restaino, R.; and Chanussot, J. 2018. Full Scale Regression-Based Injection Coefficients for Panchromatic Sharpening. *IEEE Transactions on Image Processing*, 3418–3431.

Wald, L. 2002. Data Fusion. Definitions and Architectures - Fusion of Images of Different Spatial Resolutions. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot,Le Centre pour la Communication Scientifique Directe - HAL - Diderot*.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6): 691–699.

Wang, Y.; Deng, L.-J.; Zhang, T.-J.; and Wu, X. 2021. SSconv: Explicit Spectral-to-Spatial Convolution for Pansharpening. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, DOI: 10.1145/3474085.3475600.

Wu, Z.-C.; Huang, T.-Z.; Deng, L.-J.; Vivone, G.; Miao, J.-Q.; Hu, J.-F.; and Zhao, X.-L. 2020. A New Variational Approach Based on Proximal Deep Injection and Gradient Intensity Similarity for Spatio-Spectral Image Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 6277–6290.

Xing, Y.; Zhang, Y.; He, H.; Zhang, X.; and Zhang, Y. 2023. Pansharpening via frequency-aware fusion network with explicit similarity constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Xu, Z.-Q. J.; Zhang, Y.; Luo, T.; Xiao, Y.; and Ma, Z. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A Deep Network Architecture for Pan-Sharpening. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European Conference on Computer Vision(ECCV)*, 328–345. Springer.

Yedla, R. R.; and Dubey, S. R. 2021. On the performance of convolutional neural networks under high and low frequency information. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part III 5*, 214–224. Springer.

Zhang, T.-J.; Deng, L.-J.; Huang, T.-Z.; Chanussot, J.; and Vivone, G. 2023. A Triple-Double Convolutional Neural Network for Panchromatic Sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 9088–9101.

Zhou, H.; Liu, Q.; and Wang, Y. 2022. PanFormer: A transformer based model for pan-sharpening. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.