Hierarchical Dual-Domain Fusion with Frequency-Guided Spatial Modeling for Pan-Sharpening

Huangqimei Zheng[1], Chengyi Pan[1], Qian Jiang[1*], Wei Zhou[2], Xin Jin[1]

[1]School of Software, Yunnan University, Kunming 650000, China

[2]School of Engineering, Yunnan University, State Key Laboratory of Vegetation Structure, Function and Construction (VegLab), Kunming, 650000, China

{zhenghuangqimei, panchengyi}@stu.ynu.edu.cn, {jiangqian, zwei, xinjin}@ynu.edu.cn

Abstract

Pan-sharpening aims to generate high-resolution multispectral images by integrating the spectral richness of low-resolution multispectral images with the spatial details of high-resolution panchromatic images. Although frequency-domain modeling shows great potential in this field, most existing methods are still limited to spatial-domain processing or fail to effectively capture the contextual interactions between frequency and spatial features. To address these issues, we propose a novel multi-scale frequency-spatial collaborative fusion approach. A Frequency-Spatial U-Net is introduced as the backbone network, in which frequency-spatial modeling blocks are embedded to progressively enhance the frequency-guided spatial contextual modeling capability across layers. To this end, we design a Dual Branch Frequency Attention module that adaptively enhances high- and low-frequency information. In addition, we introduce fine-mid-coarse resolution branches and devise a main-auxiliary multi-scale reconstruction loss to facilitate collaborative optimization. The effectiveness of the proposed model is validated through extensive experiments, demonstrating superior performance in both qualitative and quantitative evaluations. Moreover, our model achieves the fastest inference time among all compared methods, striking an excellent balance between accuracy and efficiency.

Code — https://github.com/Z-HQM/MSFSNet

Introduction

Due to limitations in sensor technology, generating high-resolution multispectral images that combine the spatial resolution of panchromatic (PAN) images and the spectral fidelity of low-resolution multispectral (MS) images is the goal of pan-sharpening. As a critical preprocessing step for remote sensing image analysis, pan-sharpening plays an important role in downstream tasks such as object detection, land cover classification, and change detection (Shah, Younan, and King 2008). With the rise of deep learning, the mainstream techniques of pan-sharpening have shifted from traditional methods relying on handcrafted priors to learning-based paradigms, resulting in significant improvements in fusion capability.

*Corresponding author.

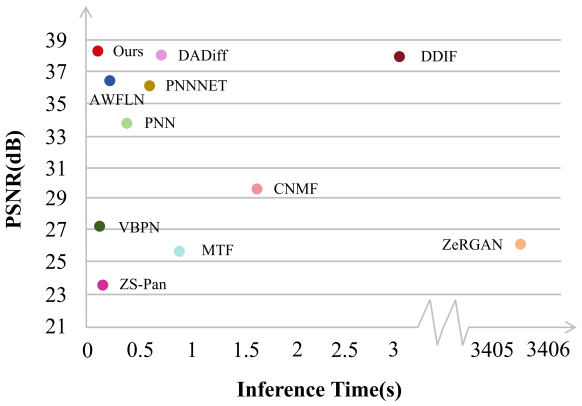

Figure 1: Inference time and PSNR for different pan-sharpening methods on the WorldView-II dataset.

Early deep models mainly adopted convolutional neural networks (CNN) to extract and fuse features from PAN and MS images (Masi et al. 2016)(Yang et al. 2017). Convolution kernels in CNN are locally frequency-selective, but due to the limited receptive field, they primarily extract local details and tend to focus on low-frequency information. As shown in Fig. 2(a), traditional CNN architectures lack explicit modeling of cross-frequency interactions and are incapable of capturing long-range dependencies. Some later studies partially compensate for these limitations of basic CNN. Encoder-decoder backbone networks, such as U-Net (Ronneberger, Fischer, and Brox 2015), alleviate the over-localization problem by fusing multi-scale features through downsampling-upsampling operations and skip connections. The Transformer architecture further expands the receptive field via the self-attention mechanism (Dosovitskiy et al. 2020)(Ma et al. 2024). However, these methods still learn the fusion mapping directly in the original spatial domain. Even when filters are used, the extracted frequency-band information remains relatively independent, lacking explicit interactions between frequencies.

Several frequency-domain modeling methods attempt to explicitly process image features in the frequency space (Tu et al. 2004)(Yu et al. 2022), as illustrated in Fig. 2(b), by applying operations on specific frequency bands. According

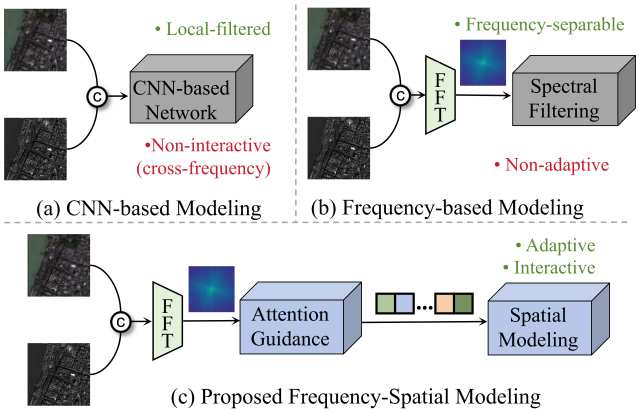

Figure 2: Comparison of Different Image Modeling Approaches.

to both the theory and practice of the Fast Fourier Transform (FFT), learning in the frequency domain is equivalent to having a global receptive field over the image, which shows advantages in preserving texture details and high-frequency information (Frigo and Johnson 1998)(Tancik et al. 2020). However, traditional frequency-domain designs are often static and non-learnable, making them incapable of adaptively emphasizing task-relevant frequencies. Furthermore, they lack spatial contextual information, which easily leads to spectral distortion or spatial misalignment. Given the complementary nature of spatial and frequency domain methods, some studies have attempted to integrate both to achieve superior fusion performance (Zhou et al. 2022b)(Tan et al. 2024). These studies demonstrate that spatial-frequency collaboration is more promising than modeling in either domain alone. However, most current dual-domain fusion networks still face significant challenges. First, most methods adopt parallel or shallow interaction structures, where spatial and frequency branches are fused through simple concatenation or weighting, lacking deep cross-domain feature exchange. Second, the frequency-domain methods often only use fixed masks, and the lack of adaptive design makes it difficult to fully exploit hierarchical structural priors and scale-dependent frequency characteristics within a unified framework.

Based on the above analysis, we propose a novel frequency-spatial collaborative fusion framework named Multi-Scale Frequency-Spatial Network (MS-FSNet). MS-FSNet is built upon the Frequency-Spatial UNet (FS-UNet) backbone by adding multi-scale branch designs. The image modeling module of FS-UNet, shown in Fig. 2(c), is implemented as the Frequency-Spatial Modeling Block (FSMB), whose core idea is to enhance spatial modeling capability through frequency-aware attention mechanisms. Specifically, we introduce a Dual Branch Frequency Attention (DBFA) module for adaptive modeling of high- and low-frequency components, and jointly employ the Spatial Sequence Mamba Modeling (SSMM) module based on an efficient state-space modeling mechanism to realize frequency-guided spatial contextual modeling. Through Mamba's unique spatial scanning and state fusion strategies, FSMB captures global sequence relations while preserving local neighborhood structures. FS-UNet adopts a UNet-style encoder-decoder structure to realize coarse-to-fine feature fusion layer-by-layer, further improving spatial detail restoration and spectral preservation in the fused image. As shown in Fig. 1, the proposed method achieves highly competitive computational efficiency while maintaining superior fusion quality. Finally, three fusion branches at different scales progressively enhance edge and texture details in a fine-mid-coarse manner.

The main contributions of this paper are summarized as follows:

- We propose a novel multi-scale dual-domain collaborative framework for pan-sharpening, MS-FSNet, and for the first time design a frequency-attention-guided state-space sequence modeling scheme, deeply coupled into the U-Net encoder-decoder architecture as FS-UNet.
- We design DBFA, which leverages FFT-based frequency analysis and dual-branch attention to adaptively enhance frequency features. Guided by DBFA, FSMB achieves frequency-aware global spatial modeling.
- Through three resolution branches, we realize multi-scale dual-domain feature integration to enhance edge and texture details. We also introduce a main-auxiliary multi-scale reconstruction loss to improve gradient complementarity between branches. Experiments on multiple public datasets demonstrate superior performance, and extensive results show that our method outperforms state-of-the-art approaches both qualitatively and quantitatively.

Related Work

DL-Based Pan-sharpening Methods

Early deep-learning methods were dominated by CNN, whose core advantage is the ability to extract local spatial features. Masi et al. (Masi et al. 2016) first proposed the three-layer CNN structure PNN, effectively fusing PAN and MS images. Subsequent methods adopted deeper network structures and multi-scale convolutional designs to improve fusion accuracy, while residual connections were introduced to enhance information propagation efficiency(Yang et al. 2017)(Wei et al. 2017)(Yuan et al. 2018).

To overcome the loss of long-range dependencies caused by the limited receptive field of convolutions, using a U-Net framework or stacking attention modules is a common practice (Peng et al. 2023)(Zheng et al. 2025). Combining channel- and spatial-attention mechanisms allows joint modeling between channel and spatial domains and also improves contextual alignment (Woo et al. 2018). Methods such as INNformer and DCTransformer introduce the Transformer into image-fusion technology to capture cross-pixel dependencies globally(Zhou et al. 2022a)(Ma et al. 2024). These studies have shown that the Transformer and its variants can outperform traditional CNN methods in low-level vision tasks; however, their window partitioning or invertible frameworks bring extra computational and engineer-

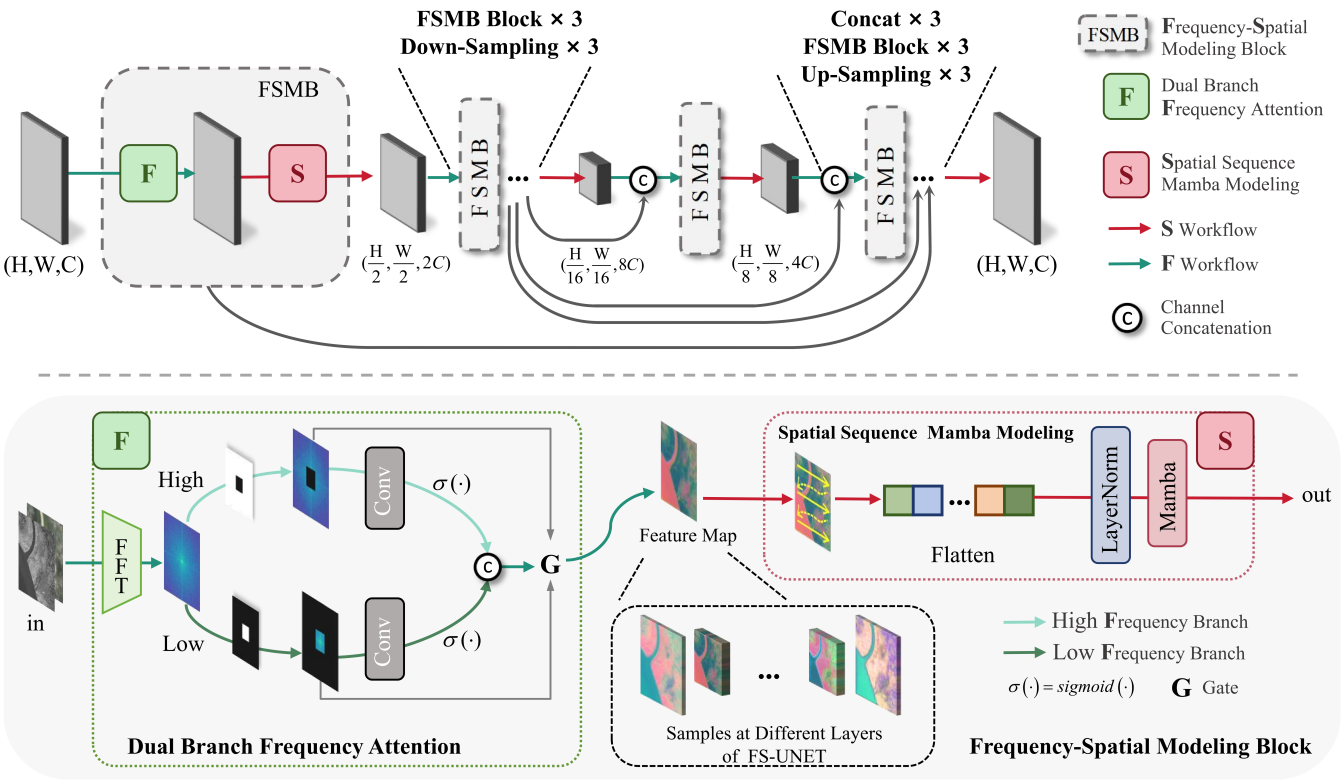

Figure 3: Overview of the FS-UNet Backbone Structure. The bottom part shows an enlarged view of the FSMB module at each layer. The FS-UNet follows the classical four-stage U-Net encoder-decoder framework. At each downsampling and upsampling stage, traditional convolution modules are replaced by FSMB to achieve deep integration of frequency enhancement and sequential modeling. During downsampling, the input features are first projected via 1×1 convolution, followed by max-pooling for spatial downsampling, and then sent to FSMB for frequency-guided modeling. Correspondingly, upsampling is performed by transposed convolution, and the skip-connection features from the encoder are concatenated by channels before being passed into FSMB. This design enables FS-UNet to effectively capture hierarchical structural information.

ing complexity. Recently, state-space models(SSM) have attracted attention for effectively alleviating the high computational complexity of existing Transformer architectures. The Mamba SSM offers linear-complexity long-range modeling capability, and there is already work applying Mamba to pan-sharpening(He et al. 2025).

Fourier Transform

The Fourier transform is a classical frequency-domain analysis tool. Its advantage in image-fusion tasks lies in the ability to distinguish high- and low-frequency components (Lasaponara and Masini 2012). In pan-sharpening tasks, several frequency-domain modeling approaches leverage the explicit separability of the Fourier domain to improve the spatial-spectral consistency of fused images. SFIIN embeds FFT into the network and adds a parallel frequency branch alongside the spatial branch (Zhou et al. 2022b). MSDDN explores the complementary information of spatial and frequency domains through a multi-scale dual-domain guidance network(He et al. 2023). HFIN proposes a new parallel multi-branch integration strategy that subdivides image information into spatial, global-Fourier, and local-Fourier branches, effectively enhancing fine-grained detail restoration(Tan et al. 2024). These studies demonstrate the broad prospects of FFT-based techniques in pan-sharpening tasks.

Proposed Method

In this section, we present the proposed MS-FSNet pan-sharpening method in detail. Our approach is a multi-scale frequency–spatial collaborative image-fusion framework whose backbone is outlined in Fig. 3, and the overall fusion strategy is illustrated in Fig. 5. We further introduce the core FSMB in FS-UNet, which guides the model to perform spatial modeling after frequency-domain attention enhancement. Within this, two subsections describe DBFA and SSMM in turn.

Frequency-Spatial Modeling Block

FSMB is the most important innovative module in FS-UNet; it aims to guide the network to perform spatial modeling after frequency-domain attention enhancement. Its structure is shown in the lower part of Fig. 3 and consists of DBFA (F) followed by SSMM (S), enabling information interaction along the way.

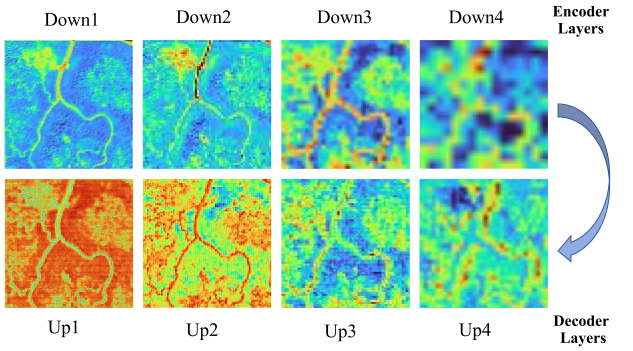

Figure 4: Visualization of Frequency-Aware Responses at Different Layers. This figure shows the importance modeling results of DBFA across multiple scales. It can be observed that the model exhibits different frequency responses to edge, texture, and structural regions at different depths.

Dual Branch Frequency Attention. The main goal of DBFA is to perform Fourier frequency-domain analysis on an input feature map and achieve frequency-selective enhancement via explicit high- and low-frequency branches. Its structure corresponds to the parts labeled F in Fig. 3. Given an input feature map $x \in \mathbb{R}^{B\times C\times H\times W}$, a 2-D orthonormal FFT is first applied (Frigo and Johnson 1998), and the magnitude spectrum is obtained:

$$X(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} x(\alpha,\beta)\, e^{-j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}, \tag{1}$$

$$\text{amp} = |X(u,v)|, \tag{2}$$

where $x(\alpha,\beta)$ denotes the spatial-domain input and $X(u,v)$ its frequency-domain representation. Binary masks are used to separate low- and high-frequency regions:

$$\text{amp}_{\text{low}} = \text{amp}\cdot M_{\text{low}}, \qquad \text{amp}_{\text{high}} = \text{amp}\cdot M_{\text{high}}, \tag{3}$$

where $M_{\text{low}} \in \{0,1\}^{H\times W}$ denotes the low-frequency mask and M_{high} the high-frequency mask. Subsequently, frequency-attention modeling is performed by feeding the processed spectral amplitudes into a 1×1 convolution module separately:

$$\text{Attn}_{\text{low}} = \sigma\big(\text{Conv}_2(\text{ReLU}(\text{Conv}_1(\text{amp}_{\text{low}})))\big), \tag{4}$$

$$\text{Attn}_{\text{high}} = \sigma\big(\text{Conv}_2(\text{ReLU}(\text{Conv}_1(\text{amp}_{\text{high}})))\big), \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Multiplying Attn_{low} and $\text{Attn}_{\text{high}}$ with the original feature map x yields x_{low} and x_{high}. After channel concatenation, an adaptive gating fusion produces the frequency-enhanced feature:

$$\text{Gate} = \sigma\big(\text{Conv}_g([x_{\text{high}}, x_{\text{low}}])\big), \tag{6}$$

$$x_{\text{fused}} = \text{Gate}\cdot x_{\text{high}} + (1-\text{Gate})\cdot x_{\text{low}}, \tag{7}$$

where $[x_{\text{high}}, x_{\text{low}}]$ denotes the channel-wise concatenation of the high- and low-frequency features. x_{fused} is a sample multi-channel feature map produced by DBFA, whose colourised visualisations at different levels of FS-UNet are shown in Fig. 4. In the deeper layers of the network, DBFA models large structures in the small-scale image, whereas in the shallow layers it enhances fine textures and edge contours.

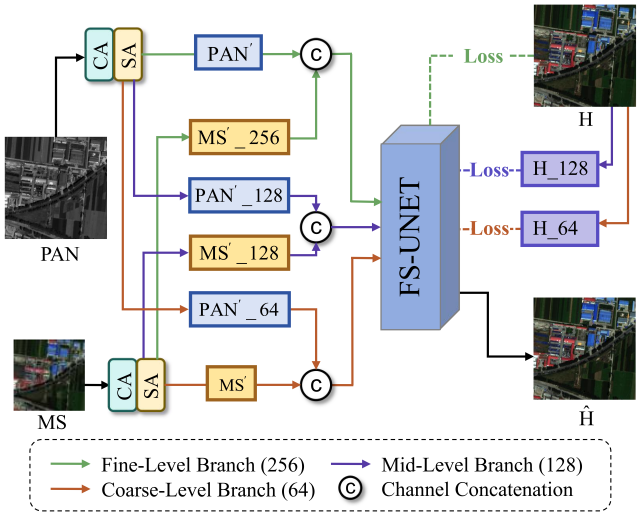

Figure 5: Multi-Scale Collaborative Fusion Strategy in MS-FSNet. The three-scale multi-branch fusion structure guides features of different resolutions into a unified FS-UNet network. At the final stage, outputs from all branches are fused and jointly optimized via loss constraints.

Spatial Sequence Mamba Modeling. SSMM is a sequence-modeling module based on Mamba whose goal is to further improve context and spatial modeling capability, as indicated by the part marked S in Fig. 3. To fit the input format required by the Mamba model, the frequency-enhanced feature map $x_{\text{fused}} \in \mathbb{R}^{B\times C\times H\times W}$ is first flattened into a sequence form:

$$x_{\text{seq}} = \text{Flatten}(x_{\text{fused}}) \in \mathbb{R}^{B\times (HW)\times C}. \tag{8}$$

where the channel vector at each spatial position is regarded as a sequence token. The sequence is then fed into a Mamba Block with a normalization layer for sequence modeling:

$$x_{\text{out}} = \text{Mamba}\big(\text{LayerNorm}(x_{\text{seq}})\big). \tag{9}$$

Mamba employs a gating mechanism to strengthen modeling of important tokens (Gu and Dao 2023), thereby enabling long-range dependency modeling and dynamic frequency-response adjustment. With this module, the network can integrate information across regions on top of the frequency-enhanced features, effectively capturing contextual structure and semantic distribution.

Overall Fusion Strategy

To fully exploit fusion capability at different scales, a three-branch multi-scale fusion structure is adopted, as shown in Fig. 5 (with input PAN resolution of 256×256 as an example). Before branching, PAN and MS inputs undergo cascaded channel and spatial attention:

$$\text{PAN}' = \text{SA}\big(\text{CA}(\text{PAN})\big), \tag{10}$$

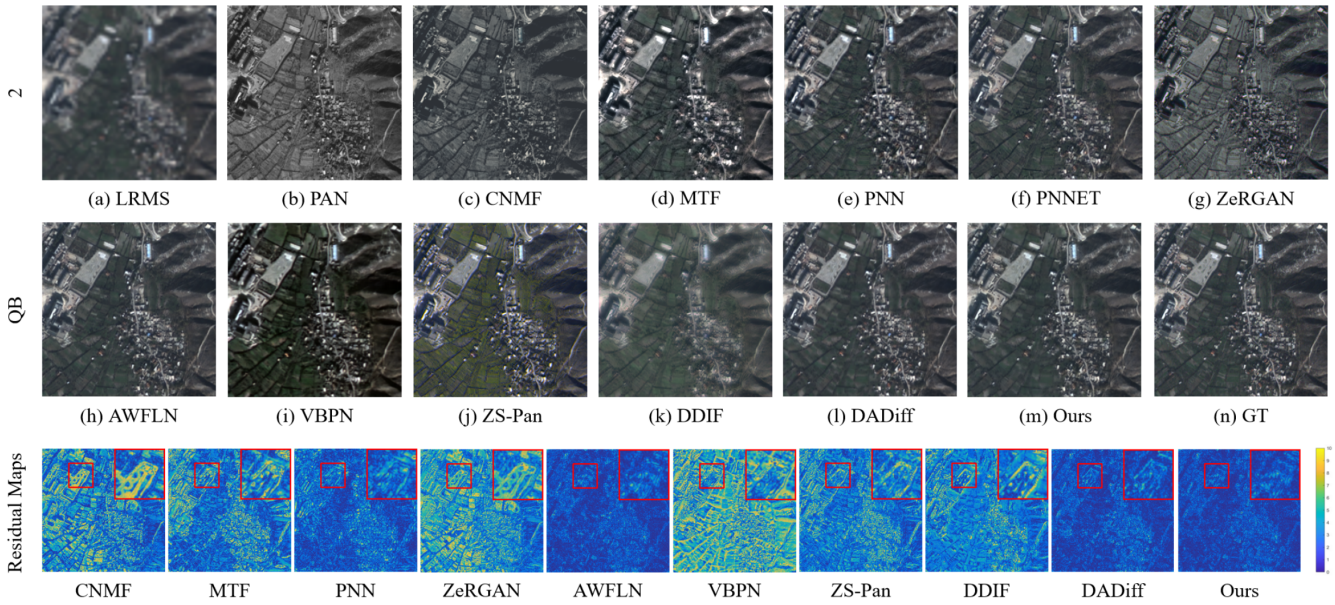

Figure 6: Qualitative comparisons on QB dataset.

$$\mathrm{MS}' = \mathrm{SA}\big(\mathrm{CA}(\mathrm{MS})\big) \tag{11}$$

where $\mathrm{CA}(\cdot)$ denotes channel attention and $\mathrm{SA}(\cdot)$ denotes spatial attention. This front-end attention strengthens feature representation before the images enter the backbone network (FS-UNet). Next, PAN' and MS' are resized to three resolutions forming three pairs of inputs for the Fine-Level Branch (256), Mid-Level Branch (128), and Coarse-Level Branch (64). Each pair is channel-concatenated at its respective resolution and fed into FS-UNet, producing outputs $\hat{H}_{256}$, $\hat{H}_{128}$, $\hat{H}_{64}$. $\hat{H}_{128}$ and $\hat{H}_{64}$ are then upsampled to 256×256 and fused along the channel dimension. The final output image is

$$\hat{H} = \mathrm{Fusion}\big([\hat{H}_{256},\ \mathrm{Up}(\hat{H}_{128}),\ \mathrm{Up}(\hat{H}_{64})]\big). \tag{12}$$

Multi-Scale Loss Function Design

To enforce consistency of the fused image at multiple scales, a composite loss is defined, consisting of a main-scale loss, auxiliary-scale losses, a color-preservation loss, and a structure-preservation loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{128} + \mathcal{L}_{64} + \mathcal{L}_{\text{color}} + \mathcal{L}_{\text{ssim}}. \tag{13}$$

where $\mathcal{L}_{\text{main}}$ is the main-scale reconstruction loss, defined as the L1 difference between the final fused image $\hat{H}$ and the ground-truth image H. The auxiliary multi-scale supervision is

$$\mathcal{L}_s = \big\|\hat{H}_s - H_s\big\|_1, \qquad s \in \{128, 64\}, \tag{14}$$

where $\hat{H}_s$ is the FS-UNet output at resolution $s \times s$ and H_s is the reference obtained by bilinear downsampling of H. This auxiliary loss stabilises training of the intermediate branches and guides the network to learn fusion capability at each scale. $\mathcal{L}_{\text{color}}$ and $\mathcal{L}_{\text{ssim}}$ denote the color-preservation term and the structural-similarity term, respectively. $\mathcal{L}_{\text{total}}$ appropriately guides the network to learn in a balanced manner across multiple scales.

Experimental Study

Datasets and Benchmarks

Our experiments utilize pairs of PAN and MS images from QuickBird (QB), WorldView-II (WV-II), and the Maryland dataset to validate our fusion model. According to the Wald protocol, we treat high-resolution multispectral (HRMS) images as the reference ground truth. By downsampling the HRMS images, we obtain low-resolution multispectral (LRMS) images for training and testing. In the three datasets, the original data are cropped into 256×256×1 patches for panchromatic (PAN) images and 64×64×3 patches for LRMS images. We compare the proposed method with 10 state-of-the-art methods, including 2 classical fusion methods and 8 representative deep learning-based methods. The two classical methods are CNMF (Baronti et al. 2011) and MTF (Vivone et al. 2013). The eight deep learning-based methods include PNN (Masi et al. 2016), PNNNET (Yang et al. 2017), ZeRGAN (Diao et al. 2022), AWFLN (Lu et al. 2023), VBPN (Zhang et al. 2024), ZS-Pan (Cao et al. 2024a), DDIF (Cao et al. 2024b), and DADiff (Zheng et al. 2025). Among them, PNN, PNNNET, AWFLN, VBPN, and ZS-Pan are CNN-based fusion methods; ZeRGAN is a GAN-based fusion method; and DDIF and DADiff are diffusion model-based fusion methods.

Training Details and Evaluation Metrics

In our experiments, the proposed model is implemented using PyTorch version 2.2.0 and runs on an NVIDIA A100 GPU. During the training phase, to ensure consistency

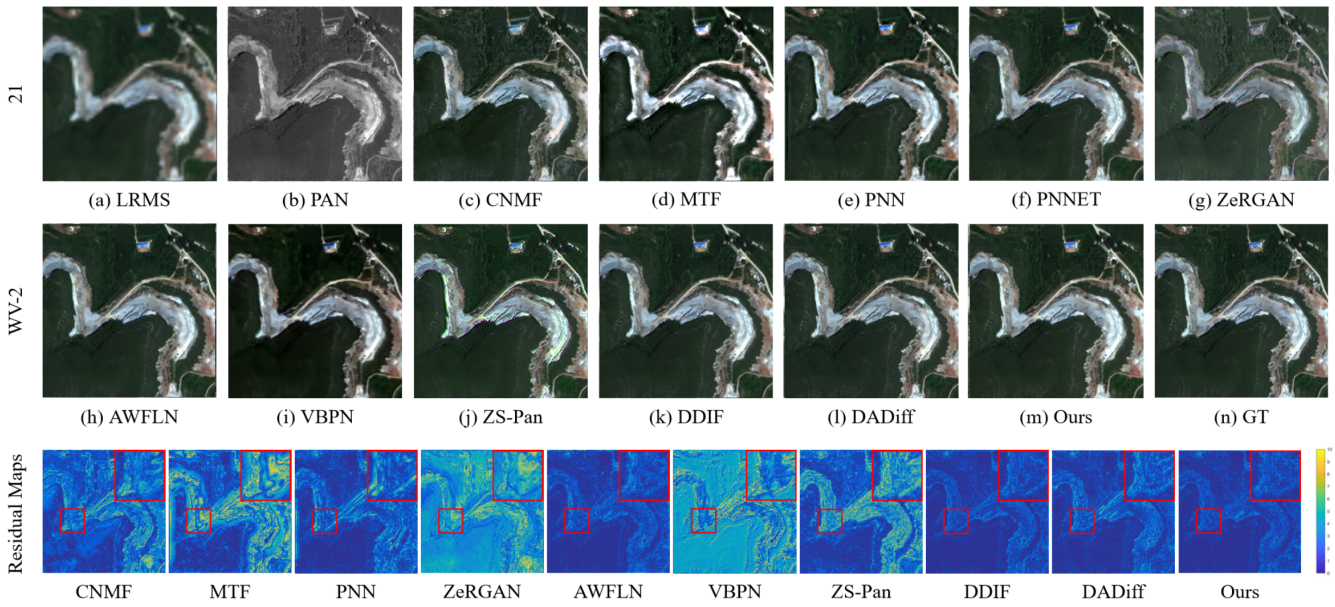

Figure 7: Qualitative comparisons on WV-2 dataset.

Datasets	Methods	PSNR↑	SAM↓	ERGAS↓	UIQI↑	CC↑	SSIM↑	D_λ↓	D_S↓	QNR↑
QB	CNMF(Baronti et al. 2011)	27.8072	0.0482	4.7697	0.7346	0.8371	0.7677	0.1415	0.1701	0.7125
	MTF(Vivone et al. 2013)	24.8132	0.0531	6.1068	0.7383	0.8594	0.7379	0.1322	0.1695	0.7207
	PNN(Masi et al. 2016)	31.1067	0.0455	3.0338	0.8932	0.9519	0.8645	0.1069	0.1643	0.7464
	PNNNET(Yang et al. 2017)	33.4401	0.0402	2.3649	0.9083	0.9655	0.8867	0.0830	0.1637	0.7669
	ZeRGAN(Diao et al. 2022)	24.5317	0.0939	6.7938	0.6542	0.7621	0.7089	0.1207	0.1683	0.7313
	AWFLN(Lu et al. 2023)	34.1836	0.0399	2.0945	0.9163	0.9672	0.9014	0.0882	0.1614	0.7648
	VBPN(Zhang et al. 2024)	23.4720	0.1183	7.6476	0.6289	0.8776	0.6248	0.1422	0.1724	0.7099
	ZS-Pan(Cao et al. 2024a)	24.6058	0.1165	6.4788	0.6389	0.7982	0.6273	0.1132	0.1677	0.7381
	DDIF(Cao et al. 2024b)	26.9089	0.0474	4.4254	0.7649	0.8957	0.7679	0.1011	0.1623	0.7530
	DADiff(Zheng et al. 2025)	34.1577	0.0416	2.1372	0.9195	0.9717	0.9065	0.0878	0.1514	0.7741
	Ours	**34.6585**	**0.0397**	**2.0559**	**0.9253**	**0.9726**	**0.9076**	**0.0822**	**0.1497**	**0.7804**
	Ideal Value	∞	0	0	1	1	1	0	0	1

Table 1: Experimental results on QB dataset. The best and second-best values are bold and underlining, respectively.

across the three datasets, we use the same set of hyperparameters. The AdamW optimizer is used to minimize the loss, with an initial learning rate of 0.0001. If the validation loss does not decrease for three consecutive epochs, the learning rate is halved. We apply Exponential Moving Average (EMA) to smooth the model parameters. Common evaluation metrics are used, including PSNR (Nezhad et al. 2016), SAM (Alparone et al. 2007), ERGAS (Wald 2000), UIQI (Wang and Bovik 2002), CC (Palsson et al. 2011), and SSIM (Wang et al. 2004), as well as unsupervised metrics for real full-resolution scenarios: D_λ, D_S, and QNR (Alparone et al. 2008).

Fusion Performance Analysis

Fig. 6 and 7 present the visualization results and residual maps for QB and WV-II, respectively. To evaluate the differences between the results and the ground truth (GT), we generate residual maps using the mean absolute error, which visualize the magnitude of the discrepancies. Brighter regions in the maps indicate greater differences. In addition,

Methods	PSNR↑	SAM↓	ERGAS↓	UIQI↑	CC↑	SSIM↑
CNMF	29.4041	0.1251	6.6642	0.7601	0.8913	0.8426
MTF	25.5591	0.1271	9.8907	0.7216	0.8625	0.7979
PNN	33.8802	0.0988	3.9587	0.8914	0.9626	0.9169
PNNNET	36.0953	0.0855	3.0668	0.9063	0.9729	0.9433
ZeRGAN	26.1622	0.1729	9.4331	0.6571	0.8016	0.7643
AWFLN	36.4797	0.0892	2.8551	0.9111	0.9760	0.9474
VBPN	27.0685	0.1820	10.4366	0.6691	0.8764	0.6575
ZS-Pan	23.5493	0.1949	13.42062	0.4933	0.6884	0.6045
DDIF	37.9292	**0.0798**	2.3691	0.9156	0.9812	0.9602
DADiff	38.1268	0.0894	2.5449	0.9219	0.9793	0.9606
Ours	**38.3307**	0.0878	**2.3656**	**0.9256**	**0.9823**	**0.9610**
Ideal Value	∞	0	0	1	1	1

Table 2: Experimental results on WV-II dataset. The best and second-best values are bold and underlining, respectively.

we selectively zoom in on local regions of the residual maps to observe detailed differences. The pan-sharpened images generated by our proposed method exhibit colors and edge details that are closer to the GT. Compared with other meth-

Methods	PSNR↑	SAM↓	ERGAS↓	UIQI↑	CC↑	SSIM↑
CNMF	23.1435	0.0814	7.4974	0.5569	0.7718	0.5948
MTF	21.7299	0.0936	8.2019	0.5981	0.7942	0.5991
PNN	26.6079	0.0868	4.7818	0.7742	0.9110	0.7225
PNNNET	28.4816	0.0747	3.9026	0.8082	0.9473	0.7601
ZeRGAN	18.2682	0.1457	13.0850	0.4176	0.3581	0.5381
AWFLN	29.6284	0.0763	3.4638	0.8290	0.9589	0.7893
VBPN	18.7416	0.1213	10.0173	0.5746	0.8453	0.5192
ZS-Pan	17.5304	0.2800	12.3261	0.4538	0.6778	0.4098
DDIF	28.5376	0.0743	3.7843	0.7951	0.9441	0.7364
DADiff	29.9059	0.0741	**3.0871**	0.8277	0.9579	0.8067
Ours	**30.1913**	**0.0737**	3.2503	**0.8376**	**0.9633**	**0.8181**
Ideal Value	∞	0	0	1	1	1

Table 3: Experimental results on Maryland dataset. The best values are bold and the second-best values are underlining.

Config	DBFA	SSMM	3-B	PSNR↑	SAM↓	ERGAS↓	SSIM↑
(I)	✗	✓	✓	33.6824	0.0489	2.4768	0.8843
(II)	✓	✗	✓	33.7947	0.0473	2.4236	0.8890
(III)	✓	✓	✗	34.3318	0.0415	2.1512	0.9015
Ours	✓	✓	✓	**34.6585**	**0.0397**	**2.0559**	**0.9076**

Table 4: Ablation study results on the testset of QB. Bold indicates the best value.

ods, our results show darker tones, fewer bright spots, and the smallest differences in both spatial and spectral aspects. This further demonstrates the superiority of our method.

Compared with other existing methods, our proposed approach shows greater advantages in quantitative metrics, as shown in Tab. 1 to 3. Our method outperforms all current methods on all evaluation metrics for the QB dataset. On the WV-2 and Maryland datasets, our method achieved the best results in most metrics. These results demonstrate that our method significantly outperforms other deep learning-based algorithms, indicating that the fused images generated by our model exhibit better texture details and overall image quality.

Ablation Studies

To verify the effectiveness of the proposed design, we conducted ablation experiments on the QB dataset, focusing on three key components: the DBFA module, the SSMM module, and the multi-scale three-branch fusion strategy (3-B). As shown in Tab. 4, removing DBFA leads to performance degradation due to the loss of frequency-aware capability, while replacing SSMM with lightweight convolutions weakens global dependency modeling and spatial-spectral consistency. Additionally, removing the mid- and low-resolution branches from the 3-B structure results in declines in both PSNR and ERGAS, indicating the importance of multi-scale guidance. These results collectively demonstrate that all three components contribute significantly to the final fusion quality.

Computational Complexity Discussion

To evaluate the efficiency of our method, we compare model parameters, inference time, and memory usage across methods (Tab. 5). Our model, with 14.38M parameters, achieves the fastest inference and comparatively low memory usage (1042MB).

Methods	Publication	Params(M)	Test Time(s)
CNMF	JSTSP	-	1.657
MTF	GRSL	-	0.823
PNN	Remote Sensing	**0.08**	0.475
PNNNET	ICCV	0.16	0.551
ZeRGAN	TNNLS	0.92	3405.695
AWFLN	TGRS	0.16	0.215
VBPN	TNNLS	2.86	0.111
ZS-Pan	Information Fusion	0.76	0.166
DDIF	Information Fusion	20.50	3.046
DADiff	Information Fusion	15.27	0.633
Ours	-	14.38	**0.101**

Table 5: Average test time and Parameters of Different Methods.

Datasets	Methods	PSNR↑	SAM↓	ERGAS↓	UIQI↑	CC↑	SSIM↑
QB	PNN	28.7255	0.0499	3.9627	0.8829	0.9375	0.8050
	PNNNET	31.1195	0.0454	3.8343	0.8921	0.9462	0.8265
	ZeRGAN	22.1911	0.1021	8.6886	0.5986	0.7315	0.6538
	AWFLN	33.2943	0.0412	2.2492	0.9158	0.9597	0.8781
	VBPN	20.4039	0.1210	8.8806	0.5770	0.8560	0.5870
	ZS-Pan	21.9894	0.1386	8.2083	0.5671	0.7417	0.5338
	DDIF	24.0690	0.0497	5.3070	0.7273	0.8751	0.7258
	DADiff	33.1046	0.0401	2.1989	0.9144	0.9625	0.8852
	Ours	**33.6789**	**0.0399**	**2.1408**	**0.9215**	**0.9657**	**0.8926**
Maryland	PNN	24.1067	0.0933	5.7519	0.6171	0.8930	0.6996
	PNNNET	26.5757	0.0827	4.6798	0.7269	0.8943	0.7035
	ZeRGAN	16.6148	0.1791	14.8351	0.3679	0.3018	0.4808
	AWFLN	28.7239	0.0795	3.9223	0.7483	0.9238	0.7347
	VBPN	16.3473	0.1327	11.6702	0.5042	0.7994	0.4563
	ZS-Pan	15.6004	0.3337	13.2709	0.3025	0.5367	0.3737
	DDIF	25.9891	0.0768	4.8558	0.7192	0.9179	0.6689
	DADiff	28.6291	0.0757	3.7327	0.7498	**0.9332**	0.7591
	Ours	**28.8839**	**0.0742**	**3.4530**	**0.7541**	0.9293	**0.7613**

Table 6: Experimental results on generalizability on different datasets.

Generalizability Experiments

To evaluate generalization, we conduct cross-dataset tests between the Maryland and QB datasets, as illustrated in Tab. 6. Despite some performance drop, our method consistently outperforms others with the least degradation.

Conclusion

In this work, we propose a multiscale frequency–spatial collaborative fusion network for pan-sharpening, which establishes a deep fusion pathway between the frequency and spatial domains. By integrating adaptive frequency attention and Mamba-based spatial modeling into a UNet-based architecture, the network captures global cross-frequency and cross-scale dependencies while preserving local structures. Multi-scale branching further enables progressive detail enhancement. Experiments on QuickBird, WorldView-II, and Maryland datasets demonstrate that MS-FSNet outperforms mainstream methods in most categories with superior spatial resolution. Future work will focus on lightweight modeling and deeper adaptive cross-domain interactions.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Nos. 62261060), Yunnan Fundamental Research Projects (Nos. 202503AG380006, 202301AW070007, 202301AU070210, 202401AT070470), and Yunnan Province Expert Workstations (202305AF150078), Yunnan Province Special Project (Grant No.202403AP140021), The Fifth Engineering Master's Degree Practice and Innovation Project of Yunnan University (Grant No. ZC-252511626), and Xingdian Talent Project in Yunnan Province of China.

References

Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; and Selva, M. 2008. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing*, 74(2): 193–200.

Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; and Bruce, L. M. 2007. Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3012–3021.

Baronti, S.; Aiazzi, B.; Selva, M.; Garzelli, A.; and Alparone, L. 2011. A theoretical analysis of the effects of aliasing and misregistration on pansharpened imagery. *IEEE Journal of Selected Topics in Signal Processing*, 5(3): 446–453.

Cao, Q.; Deng, L.-J.; Wang, W.; Hou, J.; and Vivone, G. 2024a. Zero-shot semi-supervised learning for pansharpening. *Information Fusion*, 101: 102001.

Cao, Z.; Cao, S.; Deng, L.-J.; Wu, X.; Hou, J.; and Vivone, G. 2024b. Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images. *Information Fusion*, 104: 102158.

Diao, W.; Zhang, F.; Sun, J.; Xing, Y.; Zhang, K.; and Bruzzone, L. 2022. ZeRGAN: Zero-reference GAN for fusion of multispectral and panchromatic images. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8195–8209.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Frigo, M.; and Johnson, S. G. 1998. FFTW: An adaptive software architecture for the FFT. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, 1381–1384. IEEE.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

He, X.; Cao, K.; Zhang, J.; Yan, K.; Wang, Y.; Li, R.; Xie, C.; Hong, D.; and Zhou, M. 2025. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 115: 102779.

He, X.; Yan, K.; Zhang, J.; Li, R.; Xie, C.; Zhou, M.; and Hong, D. 2023. Multiscale dual-domain guidance network for pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Lasaponara, R.; and Masini, N. 2012. Pan-sharpening techniques to enhance archaeological marks: an overview. *Satellite Remote Sensing: a new tool for Archaeology*, 87–109.

Lu, H.; Yang, Y.; Huang, S.; Chen, X.; Chi, B.; Liu, A.; and Tu, W. 2023. AWFLN: An adaptive weighted feature learning network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Ma, Q.; Jiang, J.; Liu, X.; and Ma, J. 2024. Reciprocal transformer for hyperspectral and multispectral image fusion. *Information Fusion*, 104: 102148.

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

Nezhad, Z. H.; Karami, A.; Heylen, R.; and Scheunders, P. 2016. Fusion of hyperspectral and multispectral images using spectral unmixing and sparse coding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6): 2377–2389.

Palsson, F.; Sveinsson, J. R.; Benediktsson, J. A.; and Aanaes, H. 2011. Classification of pansharpened urban satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1): 281–297.

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*, 3219–3227.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Shah, V. P.; Younan, N. H.; and King, R. L. 2008. An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets. *IEEE transactions on geoscience and remote sensing*, 46(5): 1323–1335.

Tan, J.; Huang, J.; Zheng, N.; Zhou, M.; Yan, K.; Hong, D.; and Zhao, F. 2024. Revisiting Spatial-Frequency Information Integration from a Hierarchical Perspective for Panchromatic and Multi-Spectral Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25922–25931.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33: 7537–7547.

Tu, T.-M.; Huang, P. S.; Hung, C.-L.; and Chang, C.-P. 2004. A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Geoscience and Remote sensing letters*, 1(4): 309–312.

Vivone, G.; Restaino, R.; Dalla Mura, M.; Licciardi, G.; and Chanussot, J. 2013. Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 11(5): 930–934.

Wald, L. 2000. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, 99–103. SEE/URISCA.

Wang, Z.; and Bovik, A. C. 2002. A universal image quality index. *IEEE signal processing letters*, 9(3): 81–84.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, Y.; Yuan, Q.; Shen, H.; and Zhang, L. 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE Geoscience and Remote Sensing Letters*, 14(10): 1795–1799.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A deep network architecture for pansharpening. In *Proceedings of the IEEE international conference on computer vision*, 5449–5457.

Yu, H.; Huang, J.; Zhao, F.; Gu, J.; Loy, C. C.; Meng, D.; Li, C.; et al. 2022. Deep fourier up-sampling. *Advances in Neural Information Processing Systems*, 35: 22995–23008.

Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; and Zhang, L. 2018. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3): 978–989.

Zhang, Z.; Li, H.; Ke, C.; Chen, J.; and Tian, X. 2024. Deep Variational Network for Blind Pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*.

Zheng, H.; Pan, C.; Jin, X.; Wozniak, M.; Wang, P.; Lee, S.-J.; and Jiang, Q. 2025. A pan-sharpening model using dual-branch attention-guided diffusion networks. *Information Fusion*, 120: 103076.

Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; and Liu, A. 2022a. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3553–3561.

Zhou, M.; Huang, J.; Yan, K.; Yu, H.; Fu, X.; Liu, A.; Wei, X.; and Zhao, F. 2022b. Spatial-frequency domain information integration for pan-sharpening. In *European conference on computer vision*, 274–291. Springer.