

MSPF-Net: Multi-Scale Spatial-Spectral Feature Fusion Network for Pansharpening

Wanchun Sun

sunwanchun@jsut.edu.cn

Jiangsu University of Technology

Xinyu Duan

Jiangsu University of Technology

Shujia Li

Hefei Normal University

Hui Ma

Tiangong University

Research Article

Keywords: Pansharpening, multi-scale feature fusion, high-frequency enhancement, spectral-spatial attention fusion

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11257898/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MSPF-Net: Multi-Scale Spatial-Spectral Feature Fusion Network for Pansharpening

Abstract—Pansharpening seeks to reconstruct high-spatial-resolution multispectral (MS) images by integrating the complementary spectral information of low-spatial-resolution MS images and the spatial details of high-spatial-resolution panchromatic (PAN) images. Despite their complementary nature, the substantial discrepancy in spatial resolution and information representation between the two modalities makes it difficult to simultaneously recover fine spatial structures and preserve spectral fidelity. To address this challenge, we propose a Multi-Scale Spatial-Spectral Feature Fusion Network (MSPF-Net) for effective spatial-spectral information integration. MSPF-Net establishes a hierarchical fusion framework that jointly exploits multiscale spatial structures, PAN-derived high-frequency information, and adaptive spatial-spectral interactions. Multi-scale Pyramid Feature Extraction captures structural information across different receptive fields, while PAN High-frequency Enhancement module provides explicit high-frequency spatial priors for detail reconstruction. A spectral-spatial attention fusion then adaptively regulates the contribution of complementary features to achieve a more effective balance between spatial enhancement and spectral preservation. Extensive experiments on the IKONOS, WorldView-3, and Pléiades datasets demonstrate that MSPF-Net consistently achieves strong performance under the reduced-resolution (RR) protocol, obtaining the best PSNR, RMSE, SAM, and ERGAS results across all three datasets. Full-resolution (FR) experiments further demonstrate its effectiveness under realistic full-resolution conditions. These results validate the capability of MSPF-Net to effectively reconcile spatial detail restoration and spectral fidelity, providing an effective framework for high-quality remote sensing image pansharpening.

Index Terms—Pansharpening, multi-scale feature fusion, high-frequency enhancement, spectral-spatial attention fusion.

I. INTRODUCTION

THE rapid development of remote sensing technology has created an increasing demand for high-resolution Earth observation data in applications such as land-cover classification, resource exploration, and disaster monitoring. However, sensing satellite imaging systems are constrained by sensor hardware limitations and transmission bandwidth, making it difficult to simultaneously acquire observations with high spatial and spectral resolutions. Pansharpening has therefore become an important technique in remote sensing image processing, aiming to generate high-resolution multispectral (HRMS) images by integrating the spatial details of high-spatial-resolution panchromatic (PAN) images with the spectral information of low-spatial-resolution multispectral (MS) images[1, 2]. The resulting HRMS images provide enhanced spatial structures while preserving spectral characteristics, thereby improving the quality of subsequent remote sensing interpretation and analysis.

Early pansharpening methods mainly relied on model-driven strategies, including component substitution (CS), multi-resolution analysis (MRA), and variational optimization (VO). CS-

based methods achieve spatial enhancement through component transformation and replacement, providing high computational efficiency but often introducing spectral distortions[3]. MRA-based methods exploit multiscale representations, such as pyramids and wavelets, to extract and inject spatial details, generally providing better spectral fidelity but limited adaptability to complex spatial structures[1, 2]. VO-based methods formulate pansharpening as an optimization problem with data-fidelity terms and prior constraints, providing a principled framework for incorporating physical assumptions. However, their performance depends strongly on handcrafted priors and iterative optimization, which limits computational efficiency and adaptability to diverse scenes [4]. Consequently, traditional approaches remain constrained by predefined assumptions and handcrafted representations, making it difficult to achieve a satisfactory balance between spatial detail enhancement and spectral fidelity.

With the rapid development of deep learning, data-driven methods have substantially advanced pansharpening by learning nonlinear mappings between PAN and MS images. Convolutional neural networks (CNNs) have demonstrated the capability to learn effective feature representations for pansharpening and improve spatial detail reconstruction[5]. Subsequent studies have further introduced attention mechanisms, multiscale architectures, and other advanced learning strategies to enhance feature extraction and multimodal information fusion[6, 7]. Despite these advances, several challenges remain. First, PAN images contain rich high-frequency information, including edges, textures, and local structural variations, yet existing methods often rely on generic feature extraction or implicit detail injection, which may not fully exploit informative high-frequency responses. Second, remote sensing scenes exhibit substantial variations in object scale and structural complexity, making it difficult for single-scale representations to simultaneously capture fine-grained details and broader contextual information[8]. Third, PAN and MS images differ substantially in spatial resolution, spectral dimensionality, and feature distribution. Direct feature concatenation or simple fusion strategies therefore have limited ability to establish effective cross-modal interactions, which may hinder the complementary utilization of spatial and spectral information[7].

Recent studies have attempted to alleviate these limitations through multiscale modeling, attention mechanisms, cross-modal fusion, and spatial-frequency interaction. Multiscale and multi-stream architectures enhance the representation of spatial and spectral information at different levels[8], while bilateral pyramid and attention-based strategies further improve the interaction between spatial and spectral features[7, 9]. Spatial-frequency modeling provides another effective perspective for characterizing structural information and high-frequency details[10]. Meanwhile, feature interaction and adaptive fusion mechanisms have been explored to better coordinate the complementary characteristics of PAN and MS images[11, 12]. Alt-

though these approaches have achieved encouraging performance, they generally emphasize specific aspects of the fusion process. In particular, the joint modeling of informative high-frequency PAN details, multiscale spatial structures, and spectral-spatial feature relationships remains insufficiently explored. This limitation may lead to blurred textures, spectral deviations, or spatial artifacts when complex structures and heterogeneous image characteristics are encountered. Therefore, a unified fusion framework is still needed to explicitly enhance spatial information while adaptively coordinating spatial and spectral representations.

To address these challenges, we propose a Multi-Scale Spatial-Spectral Feature Fusion Network (MSPF-Net) for high-quality pansharpening. MSPF-Net establishes a hierarchical feature learning and fusion framework that progressively enhances PAN spatial information, models spatial structures across multiple scales, and adaptively integrates complementary spatial and spectral features. Specifically, a PAN High-frequency Enhancement (PHE) module is designed to explicitly strengthen informative high-frequency responses in PAN images, improving the representation of edges, textures, and local structures. A Multi-scale Pyramid Feature Extraction (MPFE) module is introduced to construct hierarchical spatial representations and capture structural information across different scales. Furthermore, a Spectral-Spatial Attention Fusion (SSAF) module is developed to adaptively model the relationships between spatial and spectral features, facilitating effective cross-modal information interaction and promoting a better balance between spatial detail restoration and spectral fidelity.

The main contributions of this work are summarized as follows:

1) A Multi-Scale Spatial-Spectral Feature Fusion Network (MSPF-Net) is proposed for pansharpening. The network establishes a hierarchical fusion process that jointly considers high-frequency spatial enhancement, multiscale feature representation, and spectral-spatial feature interaction, enabling more effective spatial-spectral information integration.

2) A PAN High-frequency Enhancement (PHE) module is developed to explicitly enhance informative high-frequency responses in PAN images, improving the network's ability to characterize edges, textures, and local structural details.

3) A Multi-scale Pyramid Feature Extraction (MPFE) module and a Spectral-Spatial Attention Fusion (SSAF) module are introduced to model spatial structures at multiple scales and adaptively integrate complementary spatial and spectral features, respectively, thereby facilitating coordinated reconstruction of spatial details and spectral information.

4) Extensive experiments are conducted on the IKONOS, Pléiades, and WorldView-3 (WV3) datasets under both reduced-resolution (RR) and full-resolution (FR) protocols. Quantitative evaluations, visual comparisons, and ablation studies demonstrate the effectiveness and robustness of the proposed MSPF-Net.

II. RELATED WORKS

A. Traditional and Deep Learning-based Pansharpening Methods

Pansharpening aims to generate high-resolution multispectral

(HRMS) images by fusing high-spatial-resolution panchromatic (PAN) images with low-spatial-resolution multispectral (MS) images, while simultaneously preserving spatial details and spectral fidelity. Traditional pansharpening methods mainly include component substitution (CS) and MRA approaches. CS-based methods transform MS images into a specific feature space and replace or adjust the spatial component using PAN information. In contrast, MRA-based methods extract multi-scale spatial details from PAN images and inject them into MS images. Although these approaches are computationally efficient and easy to implement, their performance heavily depends on manually designed transformation models and detail injection strategies. Consequently, they often fail to characterize the complex nonlinear relationship between PAN and MS images, resulting in spectral distortion and inaccurate spatial enhancement.

With the rapid development of deep learning, CNN-based pansharpening methods have gradually become dominant due to their powerful feature representation capability. Unlike traditional approaches relying on handcrafted rules, deep neural networks can establish nonlinear mappings between PAN spatial information and MS spectral information through end-to-end learning. However, existing CNN-based methods still suffer from limited receptive fields and insufficient feature representation ability, which may lead to incomplete spatial detail recovery and weak structural modeling in complex scenes[13]. Moreover, many existing methods extract PAN and MS features separately and subsequently perform feature concatenation or progressive fusion, which cannot fully exploit the complementary characteristics between different modalities[14].

Recently, increasing attention has been paid to more effective cross-modal information interaction strategies. DUCD explicitly decomposes PAN and MS features into common and unique components through a deep unfolding framework, improving the utilization of complementary information between different modalities[14]. PDAF further explores heterogeneous complementary information between PAN and MS images by introducing local-global adaptive modulation, enabling more flexible feature fusion[15]. These studies demonstrate that fully exploiting the complementary relationship between spatial and spectral information is essential for improving pansharpening performance.

Despite the significant progress achieved by deep learning-based methods, several challenges remain unresolved, including effective spatial detail reconstruction, multi-scale structural representation, and fine-grained spatial-spectral interaction. Therefore, developing more powerful feature representation and fusion strategies remains an important research direction in pansharpening.

B. Spatial Feature Enhancement and Multi-scale Representation for Pansharpening

PAN images contain abundant high-frequency information, including edges, textures, and structural details, due to their higher spatial resolution. Therefore, effectively extracting and utilizing spatial information from PAN images is critical for improving pansharpening performance. However, high-frequency components may contain both meaningful structural information and noise. Directly injecting these components into MS

images may introduce artifacts, excessive sharpening, or spectral inconsistency. Therefore, designing adaptive spatial enhancement strategies has become an important research topic.

To enhance spatial details, recent studies have explored various strategies based on pixel differences, gradient information, and frequency-domain modeling. Pixel Difference Enhancement methods utilize pixel-level variations and multi-scale convolution operations to enhance edge and texture representations, improving the capability of networks to capture local spatial structures [16]. RFDGNet further introduces frequency decomposition and gradient-adaptive high-frequency enhancement to model spatial details more effectively while suppressing irrelevant high-frequency disturbances [17]. In addition, frequency-domain feature learning methods have been investigated to explicitly characterize the relationship between high-frequency PAN components and MS features, further improving spatial reconstruction capability.

Besides spatial detail enhancement, remote sensing images usually contain complex objects with significant scale variations, making single-scale feature representation insufficient for capturing diverse spatial structures. Consequently, multi-scale representation learning has been widely adopted in pansharpening networks. MMFN constructs a multi-scale multi-stream architecture to extract spatial and spectral features at different levels, thereby improving the representation ability of complex structures [8]. A bilateral pyramid network further establishes complementary spatial and spectral branches to achieve hierarchical feature extraction and cross-scale information interaction [7]. Furthermore, some recent studies incorporate attention mechanisms into multi-scale frameworks, allowing networks to adaptively select important features from different scales [11].

Although these methods have achieved promising performance, existing spatial enhancement and multi-scale strategies still have several limitations. On the one hand, some approaches rely heavily on explicit high-frequency extraction or local enhancement operations, which may not adapt well to diverse remote sensing scenes. On the other hand, multi-scale features are commonly fused through fixed operations such as concatenation or summation, while the dynamic interactions among different scales remain insufficiently explored.

Therefore, effectively enhancing spatial details while establishing adaptive multi-scale spatial representations remains an important challenge in pansharpening.

C. Spectral-Spatial Feature Fusion and Attention Mechanisms

Due to the significant differences between PAN and MS images in spatial resolution, spectral dimensions, and feature distributions, effective spatial-spectral information interaction is a fundamental issue in pansharpening. PAN images provide rich spatial structures but lack spectral diversity, whereas MS images contain abundant spectral information but suffer from limited spatial resolution. Therefore, an ideal fusion strategy should fully exploit their complementary characteristics while avoiding spectral distortion caused by spatial detail injection.

Recently, attention mechanisms have been widely introduced into pansharpening networks to adaptively learn the importance of different features. For example, the Residual Double-Attention Network combines channel and spatial attention mechanisms to recalibrate feature representations, enhancing spatial

detail recovery while reducing spectral distortion [13]. SSA-MRN further introduces spectral-spatial attention guidance to establish correlations among different feature dimensions, enabling more effective spatial and spectral feature modeling [9]. These studies demonstrate the effectiveness of attention mechanisms in balancing heterogeneous information from different modalities.

Beyond conventional attention mechanisms, cross-modal feature interaction strategies have also attracted increasing interest. CMIAN introduces a cross-modal information aggregation mechanism to adaptively integrate PAN and MS features according to their differences, improving complementary feature utilization [12]. FIAFN further develops feature interaction attention and adaptive feature selection strategies, enabling dynamic coordination between local PAN textures and global MS spectral representations [11]. MSSDIBNet adopts a spatial-spectral dual injection strategy to enhance spatial details while correcting potential spectral deviations, achieving a better balance between spatial enhancement and spectral preservation [18].

However, existing attention-based and cross-modal fusion methods still have limitations. First, many approaches perform spatial-spectral interaction at a single scale or feature level, limiting their ability to exploit hierarchical spatial structures. Second, existing fusion strategies mainly focus on channel recalibration or spatial weighting, while the fine-grained and dynamic interaction between multi-scale spatial information and spectral characteristics remains insufficiently investigated.

Therefore, developing an effective adaptive interaction mechanism between multi-scale spatial features and spectral information remains a key challenge for high-quality pansharpening.

III. PROPOSED ALGORITHM

A. Overall structure

The objective of pansharpening is to reconstruct high-resolution multispectral (HRMS) images from low-resolution multispectral (LRMS) and high-resolution panchromatic (PAN) images while simultaneously improving spatial resolution and preserving spectral fidelity. Due to the substantial differences between PAN and LRMS images in spatial resolution, spectral dimensionality, and feature distributions, effectively extracting spatial details from PAN, preserving spectral information from LRMS, and establishing cross-scale interactions between them are crucial for high-quality pansharpening. To address the insufficient spatial detail recovery and limited spatial-spectral interaction in existing methods, we propose a Multi-Scale Spatial-Spectral Feature Fusion Network (MSPF-Net). The proposed network consists of three core modules: the PAN High-frequency Enhancement (PHE) module, the Multi-scale Pyramid Feature Extraction (MPFE) module, and the Spectral-Spatial Attention Fusion (SSAF) module. Given the LRMS image, PAN image, high-resolution RGB image (SRGB), and low-resolution RGB image (LRGB), MSPF-Net jointly models the complementary information from these inputs to reconstruct the HRMS image. The overall architecture of the proposed network is illustrated in Fig. 1.

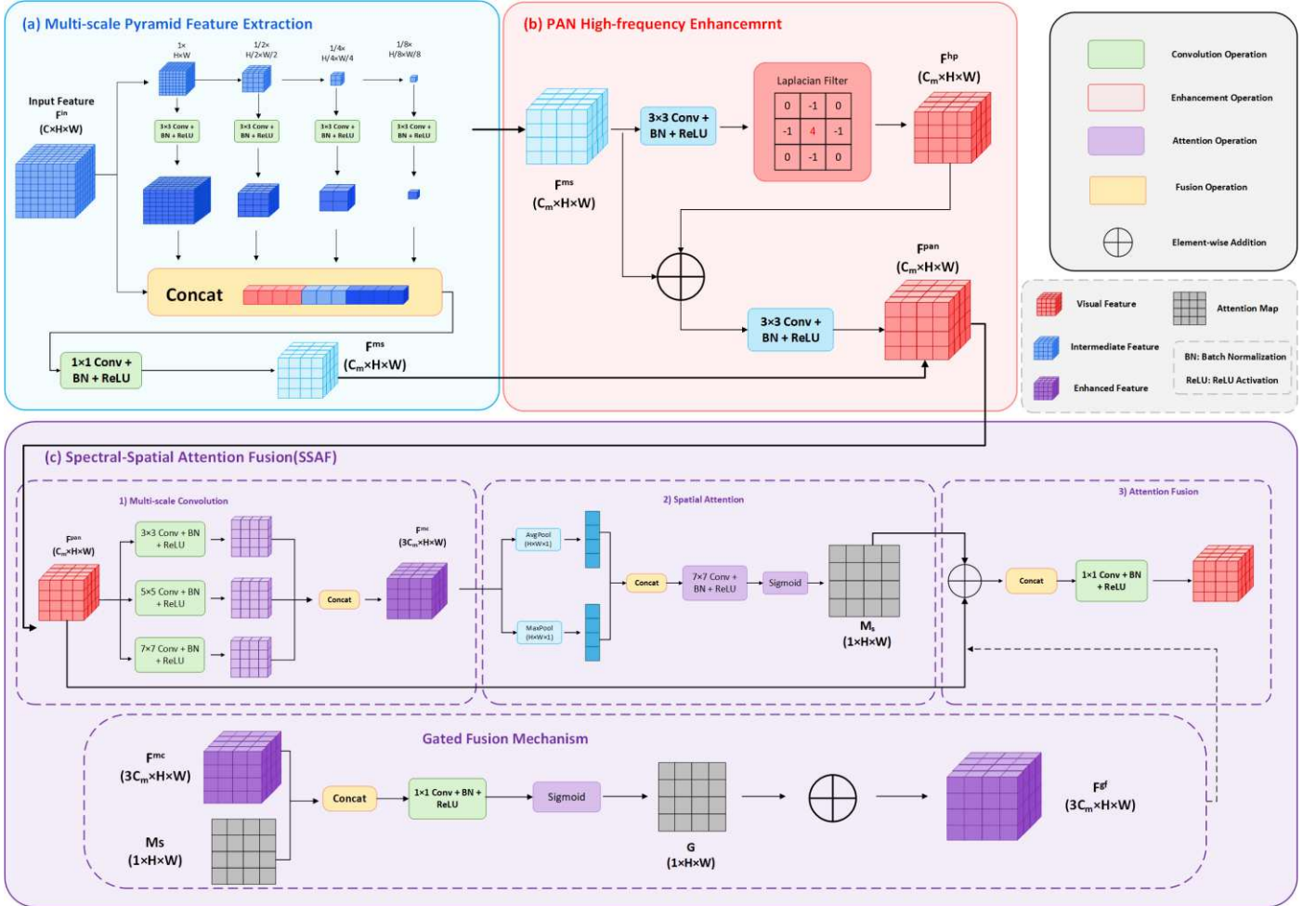


Fig. 1. Overall architecture of the proposed MSPF-Net. Given LRMS, PAN, SRGB, and LRGB inputs, the network progressively performs multi-source feature encoding, PAN high-frequency enhancement, multi-scale feature extraction, and spectral-spatial attention fusion to reconstruct the HRMS image.

For the input LRMS, PAN, SRGB, and LRGB images, separate feature extraction branches are first employed for preliminary encoding. Since the spatial resolutions of LRMS and LRGB are lower than that of the target HRMS image, a transposed convolution with a stride of 4 is first applied to upsample their spatial dimensions. For the LRMS branch, the upsampled feature is additionally preserved through a skip connection and added to the reconstructed output at the end of the network to retain the original multispectral information. For SRGB and LRGB, convolutional and upsampling-convolutional blocks are respectively employed to extract RGB-related features, which are subsequently concatenated and further transformed to construct an RGB-guided feature representation.

Specifically, let LRMS, PAN, SRGB, and LRGB be denoted by X_L , X_P , X_S , and X_R , respectively. The LRMS branch first produces:

$$F_L = U(X_L)$$

where $U(\cdot)$ denotes the spatial upsampling operation. Meanwhile, SRGB and LRGB are processed by separate convolutional feature extraction branches:

$$F_S = C_S(X_S), F_R = U(C_R(X_R))$$

where $C(\cdot)$ denotes convolution-based feature extraction. The resulting RGB features are then concatenated and further

integrated through a convolutional layer:

$$F_{RGB} = C_F([F_S, F_R])$$

On the other hand, the PAN image contains abundant high-frequency spatial information, including edges, textures, and structural details. To exploit this information more effectively, the PAN feature is processed by the PAN High-frequency Enhancement (PHE) module before entering the main fusion stage:

$$F_P^H = H(F_P)$$

where $H(\cdot)$ denotes the PAN high-frequency enhancement operation. The enhanced PAN feature is then concatenated with the upsampled LRMS feature and processed by a convolutional layer to obtain the initial fused representation:

$$F_F = C_{Fus}([F_L^H, F_P^H])$$

where F_L^H denotes the upsampled LRMS feature.

The resulting initial fused feature is subsequently fed into the Multi-scale Pyramid Feature Extraction (MPFE) module. MPFE performs parallel feature modeling at four spatial scales, including the original resolution and $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ downsampled resolutions. This design enlarges the effective receptive field while simultaneously capturing local textures and broader

spatial structures. Features extracted at different scales are restored to a unified spatial resolution and then concatenated and fused through convolution to produce the multi-scale enhanced representation:

$$F_M = P(F_F)$$

where $P(\cdot)$ denotes the multi-scale pyramid feature extraction operation.

The multi-scale enhanced feature is further refined by two cascaded Spectral-Spatial Attention Fusion (SSAF) modules. The first SSAF module integrates the multi-scale feature representation with the RGB-guided feature to enhance the complementary information among different modalities. The second SSAF module further introduces the enhanced PAN feature, progressively strengthening the interaction between spatial details and spectral-related information. The process can be formulated as:

$$F_{A1} = A_1(F_M, F_{RGB}), F_{A2} = A_2(F_{A1}, F_P^H)$$

where $A_1(\cdot)$ and $A_2(\cdot)$ denote the two cascaded attention residual fusion modules, respectively.

Finally, the refined feature is mapped to the target number of multispectral channels through a 3×3 convolution to obtain the reconstructed feature F_R . To reduce the loss of original spectral information during deep reconstruction, the upsampled LRMS input is introduced as a global residual branch and added element-wise to the reconstructed feature. The final HRMS output is therefore obtained as:

$$\hat{Y} = C_R(F_{A2}) + F_L$$

where $\hat{Y}$ denotes the reconstructed high-resolution multi-spectral image.

Overall, MSPF-Net establishes a complete processing pipeline consisting of multi-source feature encoding, PAN high-frequency enhancement, multi-scale feature extraction, spectral-spatial attention fusion, and residual reconstruction. The network enhances feature learning from three complementary perspectives: effective utilization of PAN spatial details, multi-scale spatial representation, and adaptive fusion of multi-source spectral-spatial features. The three core modules are the PAN High-Frequency Enhancement (PHE) module, Multi-scale Pyramid Feature Extraction (MPFE) module, and Spectral-Spatial Attention Fusion (SSAF) module.

B. Multi-scale Pyramid Feature Extraction (MPFE)

Spatial information in pansharpening exhibits strong scale dependency. The PAN image contains abundant high-frequency spatial structures, such as edges and textures, while the LRMS image mainly provides spectrally consistent but spatially degraded multispectral information. After the initial spatial-spectral feature fusion, the intermediate representation contains local textures, object boundaries, and regional contextual structures. However, these spatial patterns usually exist at different scales: fine-scale features are more effective for describing local details, whereas larger-scale representations are beneficial for capturing broader contextual information. A single-scale convolutional operation with a limited receptive field is insufficient to simultaneously model local details and large-scale structures, which may degrade spatial reconstruction performance in complex scenes.

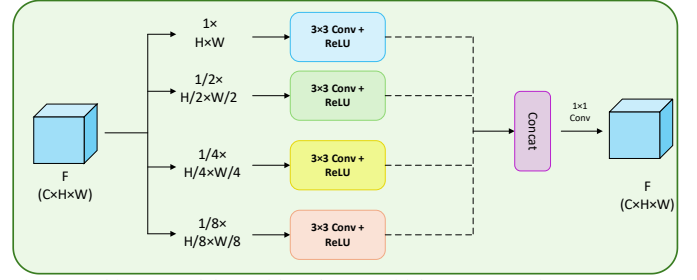


Fig. 2. Architecture of the multi-scale pyramid feature extraction (MPFE).

To address this limitation, a Multi-scale Pyramid Feature Extraction (MPFE) module is proposed, as illustrated in Fig. 2. Instead of simply stacking convolutions with different kernel sizes, MPFE constructs multi-scale feature representations by performing parallel spatial-scale transformations and subsequently aggregating them in a unified feature space. This design enables the network to jointly capture fine-grained textures and regional structural information.

Let the intermediate feature representation obtained after fusing LRMS, PAN, and RGB-related features be denoted as:

$$F \in \mathbb{R}^{C \times H \times W}$$

where C , H , and W represent the number of feature channels, height, and width, respectively. MPFE first constructs parallel feature representations at four spatial scales, including the original scale, $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ scales. The feature representation at the k -th scale can be formulated as:

$$F_k = \phi_k(D_{s_k}(F))$$

where $s_k \in \{1, 2, 4, 8\}$, $D_{s_k}(\cdot)$ denotes the average pooling operation with scale factor s_k , and $\phi_k(\cdot)$ represents a feature transformation operation consisting of a 3×3 convolution followed by a ReLU activation. For the original-scale branch, $s_1 = 1$, and thus no spatial downsampling is performed.

Specifically, the original-scale branch preserves the spatial resolution of the input feature and focuses on extracting fine-grained spatial variations:

$$F_1 = \phi_1(F)$$

For the remaining three branches, average pooling operations with kernel sizes of 2×2 , 4×4 , and 8×8 are respectively applied to reduce the spatial resolution:

$$F_2 = \phi_2(D_2(F)),$$

$$F_3 = \phi_3(D_4(F)),$$

$$F_4 = \phi_4(D_8(F)).$$

Through this parallel multi-scale construction, different branches capture spatial information with different receptive fields. The original-scale branch mainly preserves local textures and edge variations, while the lower-resolution branches aggregate contextual information over larger spatial regions. Consequently, the feature representation gradually expands from local details to regional structural information.

Since the feature maps generated at different scales have inconsistent spatial resolutions, the lower-resolution features are

restored to the original spatial size through bilinear interpolation:

$$\hat{F}_k = U(F_k), \quad k \in \{2, 3, 4\},$$

where $U(\cdot)$ denotes the bilinear upsampling operation. After scale alignment, all feature representations share the same spatial resolution of $H \times W$.

The aligned multi-scale features are then concatenated along the channel dimension:

$$F_{\text{cat}} = \text{Concat}(F_1, \hat{F}_2, \hat{F}_3, \hat{F}_4)$$

Compared with direct element-wise addition, channel-wise concatenation preserves the complementary information extracted from different spatial scales, enabling subsequent feature learning to establish correlations among various-scale representations. A 1×1 convolution is further employed to perform cross-scale feature fusion and channel compression:

$$F_{\text{fuse}} = C_{1 \times 1}(F_{\text{cat}})$$

where $C_{1 \times 1}(\cdot)$ denotes the 1×1 convolution operation for channel interaction and feature aggregation.

To reduce the disturbance caused by multi-scale transformations and preserve the original feature representation, a residual connection is introduced:

$$F_{\text{MPFE}} = F_{\text{fuse}} + F$$

The final MPFE output F_{MPFE} integrates the original spatial-spectral fusion features with contextual information extracted from different spatial scales.

From the perspective of feature representation, MPFE explicitly establishes a local-to-regional hierarchical spatial representation rather than simply increasing the network depth. The original-scale branch preserves fine-grained spatial details, while the lower-resolution branches enlarge the receptive field to capture broader contextual structures. By aligning and aggregating multi-scale features, MPFE enables complementary interaction between local details and regional spatial information.

Moreover, the residual connection preserves the original fusion feature containing LRMS-related spectral information, reducing the potential spectral distortion introduced by multi-scale spatial processing. Therefore, MPFE provides richer multi-scale spatial representations for the subsequent Spectral-Spatial Attention Fusion (SSAF) modules and forms a progressive feature learning process from PAN high-frequency enhancement to multi-scale spatial modeling and adaptive spatial-spectral fusion.

C. PAN High-frequency Enhancement (PHE)

In pansharpening, the PAN image provides abundant spatial information, including edges, textures, and local structural details, which is essential for reconstructing high-resolution multispectral images. However, PAN images lack complete spectral information, and their high-frequency responses may contain not only useful spatial details but also noise and inconsistent spatial patterns that do not correspond to the spectral characteristics of MS images. Therefore, directly introducing PAN features into the fusion process may propagate redundant or unreliable spatial responses, leading to spectral distortion.

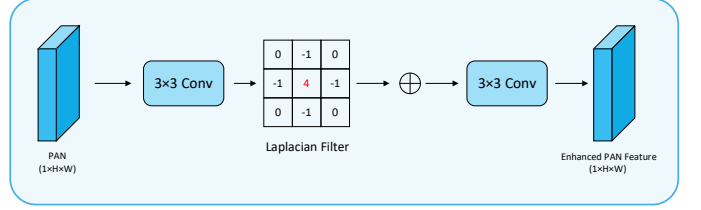


Fig. 3. Architecture of the PAN high-frequency enhancement (PHE).

To address this issue, we propose a PAN High-frequency Enhancement (PHE) module, as illustrated in Fig. 3, to explicitly enhance discriminative spatial details from PAN features before feature fusion. Instead of simply increasing the intensity of PAN responses, PHE extracts high-frequency spatial components through feature decomposition and performs adaptive refinement, enabling more effective spatial information utilization during subsequent reconstruction.

Given the input PAN image X_p , a shallow convolution layer is first applied to extract the initial PAN feature representation:

$$F_p = \phi_p(X_p)$$

Where $\phi_p(\cdot)$ denotes the initial PAN feature extraction operation. Since spatial structures can be decomposed into low-frequency components representing smooth regions and high-frequency components describing edges and textures, a high-frequency residual representation is extracted from the PAN feature:

$$F_p^H = F_p - A(F_p)$$

where $A(\cdot)$ represents the average pooling operation used to approximate the low-frequency component. The extracted high-frequency feature is then refined through convolutional transformations:

$$\hat{F}_p^H = \phi_H(F_p^H)$$

where $\phi_H(\cdot)$ denotes the high-frequency feature enhancement operation. Finally, a residual connection is introduced to preserve the original spatial representation while enhancing high-frequency details:

$$F_p^{\text{enh}} = F_p + \hat{F}_p^H$$

The residual design enables PHE to strengthen edge and texture responses while avoiding excessive modification of the original PAN spatial structure. Moreover, performing enhancement in the feature domain allows the network to learn task-oriented spatial representations rather than directly manipulating PAN pixels, improving the adaptability of spatial detail reconstruction.

After PHE, the enhanced PAN feature is fused with the upsampled LRMS feature. Let the upsampled LRMS representation be denoted as F_L^H , the initial spatial-spectral fusion feature can be formulated as:

$$F_F = C_F([F_L^H, F_p^{\text{enh}}])$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation and $C_F(\cdot)$ represents the convolutional fusion operation. In this way, the PAN branch mainly provides high-resolution spatial priors, while the LRMS branch preserves spectral-related information. The complementary information from both modalities is initially integrated before being further processed by subsequent

modules.

It should be noted that the enhanced PAN feature is not directly used for image reconstruction. Instead, it serves as an informative spatial prior for subsequent multi-scale feature modeling and spatial-spectral interaction. This strategy prevents the network from over-relying on PAN-derived details and enables MPFE and SSAF modules to further balance spatial enhancement and spectral preservation.

Therefore, PHE enhances the representation capability of high-resolution spatial information and provides more reliable spatial guidance for feature fusion. Together with Multi-scale Pyramid Feature Extraction (MPFE) and Spectral-Spatial Attention Fusion (SSAF), the proposed framework establishes a progressive feature learning process from high-frequency spatial enhancement, multi-scale structural representation, to adaptive spatial-spectral fusion.

D. Spectral-Spatial Attention Fusion (SSAF)

After PAN high-frequency enhancement and multi-scale pyramid feature extraction, the network obtains fused features with rich spatial structural information. However, high-quality pansharpening requires not only the recovery of high-resolution spatial details but also the preservation of spectral consistency. Due to the substantial differences between PAN and LRMS in imaging mechanisms, spatial resolution, and feature distributions, simple feature concatenation or element-wise fusion may not fully exploit their complementary information and can lead to an imbalance between spatial enhancement and spectral preservation.

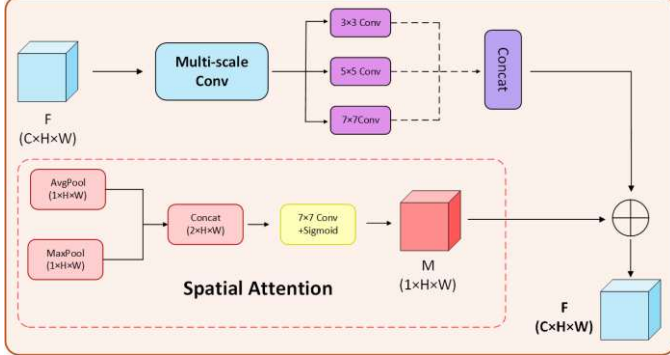


Fig. 4. Architecture of the spectral-spatial adaptive fusion (SSAF).

To further coordinate spatial information and spectral-related features, we propose a Spectral-Spatial Attention Fusion (SSAF) module, as illustrated in Fig. 4, to adaptively fuse and refine the multi-source features generated by the preceding modules. SSAF employs multi-scale convolutions to construct spatial representations with different receptive fields and introduces spatial attention to adaptively reweight feature responses at different locations. A residual connection is further adopted to preserve the basic spectral-spatial representation of the input features.

1) Multi scale Feature Interaction

Let the feature generated by MPFE be denoted as F_M , and let the RGB-guided feature be denoted as F_R . The two features are first concatenated along the channel dimension:

$$F_{\text{cat}} = [F_M, F_R]$$

The concatenated feature is then processed by three parallel convolutional branches with different receptive fields:

$$F_3 \equiv \phi_3(F_{\text{cat}}),$$

$$F_5 \equiv \phi_5(F_{\text{cat}}),$$

$$F_7 \equiv \phi_7(F_{\text{cat}}).$$

where $\phi_3(\cdot)$, $\phi_5(\cdot)$, and $\phi_7(\cdot)$ denote convolution operations with kernel sizes of 3×3 , 5×5 , and 7×7 , respectively. Thus, the multi-scale spatial features can be represented as

$$\{F_3, F_5, F_7\} \equiv \Phi(F_{\text{cat}})$$

where $\Phi(\cdot)$ denotes the multi-scale spatial feature extraction process.

The different kernel sizes provide complementary receptive fields. The 3×3 branch mainly captures local textures and edge variations, whereas the 5×5 and 7×7 branches aggregate broader contextual information. Compared with single-scale convolution, this parallel multi-scale design enables the network to capture structural variations at different spatial ranges and provides richer representations for subsequent spatial attention refinement.

2) Spatial Attention guided Feature Reweighting

Although multi-scale convolutions provide features with different receptive fields, different spatial locations do not contribute equally to HRMS reconstruction. Regions such as building boundaries, road contours, and object interfaces generally contain stronger spatial variations, whereas homogeneous regions exhibit relatively weak responses. Therefore, assigning identical weights to all spatial locations may limit the effective utilization of PAN-derived spatial information.

To address this issue, a spatial attention mechanism is introduced into the multi-scale feature representation. For an input feature F , the channel-wise average and maximum responses are first calculated as:

$$A_{\text{avg}}(F) = \frac{1}{C} \sum_{c=1}^C F_c,$$

$$A_{\text{max}}(F) = \max_c F_c.$$

Both responses have spatial dimensions of $H \times W$ and characterize the overall and salient responses across channels, respectively. They are then concatenated:

$$A(F) = [A_{\text{avg}}(F), A_{\text{max}}(F)].$$

A spatial attention map is subsequently generated by a 9×9 convolution followed by a Sigmoid activation:

$$M_s(F) \equiv \sigma(C_{9 \times 9}(A(F)))$$

where $C_{9 \times 9}(\cdot)$ denotes the 9×9 convolution and $\sigma(\cdot)$ denotes the Sigmoid function. The resulting attention map is used to adaptively reweight the input feature:

$$\tilde{F} \equiv F \square M_s(F)$$

where $\square$ denotes element-wise multiplication. In this way, spatially informative regions can receive stronger feature responses, improving the representation of edges, textures, and structural variations.

3) Cross scale Spatial Feature Refinement

To further exploit the complementary relationship among features with different receptive fields, SSAF performs progressive spatial refinement across scales. Specifically, spatial attention maps are first obtained from the 5×5 and 7×7 features:

$$M_5 = M_s(F_5), \quad M_7 = M_s(F_7)$$

The spatial response generated from the larger receptive field is then used to guide features from smaller receptive fields:

$$\tilde{F}_5 \equiv F_5 \square M_7,$$

and

$$\tilde{F}_3 \equiv F_3 \square M_5 \square M_7.$$

This progressive guidance allows broader spatial contextual information to influence local feature representations and establishes spatial correlations across different receptive fields. Compared with independently reweighting each scale, the proposed strategy promotes more effective interaction among multi-scale spatial features.

The refined 3×3 and 5×5 features are first concatenated:

$$F_{35} = [\tilde{F}_3, \tilde{F}_5],$$

and then compressed and fused using a convolution:

$$F'_{35} = \phi_{35}(F_{35}).$$

The resulting feature is further concatenated with the 7×7 feature:

$$F_{357} = [F'_{35}, F_7],$$

followed by another convolution to obtain the final fused representation:

$$F_{\text{out}} = \phi_{357}(F_{357}).$$

This hierarchical fusion process enables local, intermediate, and broader spatial information to interact progressively rather than being simply added together.

4) Residual Feature Fusion

After multi-scale spatial attention fusion, a residual connection is introduced to preserve the original spectral-spatial representation. The concatenated input feature is first projected into the target channel space using a 1×1 convolution:

$$F_{\text{res}} = C_{1 \times 1}(F_{\text{cat}})$$

The final SSAF output is then formulated as:

$$F_{\text{SSAF}} = \delta(F_{\text{out}} + F_{\text{res}})$$

where $\delta(\cdot)$ denotes the ReLU activation function.

The residual path helps preserve the original feature information during multi-scale attention fusion and facilitates stable feature propagation. More importantly, since F_{cat} contains information derived from the RGB-guided branch and the MPFE representation, the residual connection provides a direct path for preserving the underlying spectral-spatial characteristics during feature refinement.

Through multi-scale spatial modeling, spatial attention reweighting, and residual feature fusion, SSAF effectively establishes adaptive interactions between spatial details and spectral-related information, providing refined representations for HRMS reconstruction.

IV. EXPERIMENTS

A. Experimental Settings

1) *Datasets*: To evaluate the effectiveness and generalization capability of the proposed method, experiments are conducted on three widely used satellite datasets, including IKONOS, Pleiades, and WorldView-3 (WV3). These datasets cover different sensor characteristics and spectral configurations, providing a comprehensive benchmark for evaluating spatial detail reconstruction and spectral fidelity in pansharpening. Specifically, IKONOS and Pleiades contain four-band multispectral (MS) images, whereas WV3 provides eight-band MS images.

In this study, the experimental datasets are adopted from the publicly available preprocessing protocol used in previous pansharpening studies[19]. The datasets are generated according to the Wald protocol, where the original high-resolution multispectral images are regarded as reference ground truth (GT). The corresponding low-resolution multispectral (LR-MS) images are obtained through modulation transfer function (MTF)-matched filtering and spatial downsampling, while the PAN images are degraded accordingly to maintain the spatial relationship between PAN and LR-MS observations.

For the reduced-resolution (RR) setting, the processed PAN and LR-MS images have spatial sizes of 256×256 and 64×64 pixels, respectively. These paired samples provide reference images for quantitative evaluation using full-reference metrics. In addition, full-resolution (FR) experiments are performed using the original-resolution PAN and LR-MS observations. Since reference HRMS images are unavailable in this setting, the fusion quality is evaluated using no-reference metrics.

2) *Comparison methods*: To comprehensively evaluate the effectiveness of the proposed method for pansharpening, several representative approaches are selected for quantitative and qualitative comparisons. The selected methods cover different deep learning-based fusion frameworks, including CNN-based methods, generative models, and recently developed visual representation architectures, ensuring a comprehensive and representative evaluation.

Specifically, nine representative pansharpening methods are adopted for comparison, including DRPNN[20], MSDCNN[21], PanColorGAN[22], TFNet[23], PCDRN[24], AWFLN[25], LRTCFFan[26], WFANet[27], and Pan-Mamba[28].

Among them, DRPNN[20], MSDCNN[21], TFNet[23], and PCDRN[24] represent typical CNN-based pansharpening approaches, which mainly perform end-to-end feature learning to achieve information fusion between PAN and MS images. PanColorGAN[22] introduces a generative adversarial learning framework to improve the visual quality of fused results from a generative modeling perspective. AWFLN[25] and WFANet[27] further enhance the joint representation of spatial and spectral information through advanced feature interaction and adaptive fusion strategies. LRTCFFan[26] represents a recent pansharpening method that adopts more sophisticated feature modeling mechanisms. Pan-Mamba[28] is a representative state-space-model-based approach that exploits the long-range dependency modeling capability of Mamba for image fusion.

These comparison methods cover diverse technical para-

digms ranging from conventional deep convolutional architectures to emerging sequence modeling frameworks, providing a comprehensive evaluation of the proposed method.

3) *Evaluation metrics*: To quantitatively evaluate the fusion performance from different perspectives, both reference-based and no-reference quality metrics are adopted under the reduced-resolution (RR) and full-resolution (FR) settings.

For the RR experiments, where high-resolution reference multispectral images are available, six widely used reference-based metrics are employed, including Peak Signal-to-Noise Ratio (PSNR), Root Mean Square Error (RMSE), Universal Image Quality Index (UIQI), Q^{2n} , Spectral Angle Mapper (SAM), and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS). These metrics evaluate the reconstruction accuracy, spatial quality, and spectral consistency of the fused images. Specifically, higher values of PSNR, UIQI, and Q^{2n} indicate better fusion quality, while lower values of RMSE, SAM, and ERGAS correspond to smaller reconstruction errors and better spectral preservation.

For the FR experiments, since the corresponding high-resolution reference images are unavailable, three no-reference quality metrics are adopted, including spectral distortion index (D_λ), spatial distortion index (D_s), and Quality with No Reference (QNR). Among them, D_λ and D_s measure spectral and spatial distortions, respectively, while QNR provides an overall assessment of the fusion quality. Lower values of D_λ and D_s indicate reduced distortions, whereas higher values of QNR represent better overall fusion performance.

4) *Training details*: To ensure the reproducibility of the experiments, all experiments are conducted with a fixed random seed (SEED=1). The same network initialization strategy and data preprocessing protocol are adopted for all experiments. The proposed network is optimized using the Adam optimizer with an initial learning rate of 0.002 and a weight decay coefficient of 0. The learning rate is adjusted using the StepLR scheduler, where the learning rate is reduced by a factor of 0.5 every 20 epochs.

The batch size is set to 8, and the network is trained for 100

epochs in total. Model checkpoints are saved every 10 epochs during training. The proposed Model_loss is employed as the optimization objective, which constrains the reconstruction process from multiple aspects to improve the balance between spatial detail preservation and spectral fidelity.

To further reduce randomness during training, deterministic computation settings are enabled and the cuDNN benchmark mode is disabled. Unless otherwise specified, all experiments are conducted under the same training strategy and evaluation protocol to ensure fair comparisons among different methods.

B. Comparisons with the SOTA methods

To comprehensively evaluate the fusion performance of MSPF-Net, quantitative comparisons are conducted under both the reduced-resolution (RR) and full-resolution (FR) protocols using representative pansharpening methods. For the RR experiments, PSNR, RMSE, UIQI, Q^{2n} , SAM, and ERGAS are adopted, where higher values of PSNR, UIQI, and Q^{2n} and lower values of RMSE, SAM, and ERGAS indicate better performance. For the FR experiments, D_λ , D_s , and QNR are employed as no-reference quality metrics, where lower D_λ and D_s and higher QNR indicate better fusion quality.

1) Reduced-Resolution (RR) Quantitative Analysis

Tables I–III present the quantitative comparison results of different methods on the IKONOS, WorldView-3 (WV3), and Pléiades datasets under the RR protocol, respectively. Overall, MSPF-Net demonstrates strong and stable reconstruction performance across the three datasets. It consistently achieves the lowest RMSE and ERGAS, while obtaining the highest PSNR and the lowest SAM on the IKONOS and Pléiades datasets. In contrast, DASFusion shows clear advantages in UIQI and Q^{2n} and slightly outperforms MSPF-Net in PSNR and SAM on the WV3 dataset. These results indicate that different metrics capture different aspects of fusion quality, while MSPF-Net exhibits a stable advantage in reconstruction error reduction and spectral distortion suppression.

TABLE I
Quantitative Comparison of Different Methods on the IKONOS Dataset Under the RR Protocol

Methods	PSNR↑	RMSE↓	UIQI↑	Q^{2n} ↑	SAM↓	ERGAS↓
DRPNN	26.9798	13.8740	0.9273	0.8154	4.0383	3.7677
MSDCNN	27.1694	13.5968	0.9310	0.8239	4.0121	3.6587
PanColorGAN	24.0976	12.4628	0.9309	0.8224	8.4360	6.0984
TFNet	30.9652	9.1658	0.9702	0.9227	3.2528	2.4465
PCDRN	28.8464	11.3740	0.9516	0.8650	3.5955	3.0352
AWFLN	34.0407	6.0425	0.9812	0.9522	2.3388	1.6913
LRTCFPan	29.4280	9.4667	0.9598	0.9006	3.5787	2.9326
WFANet	34.0704	6.9165	0.9544	0.7832	2.3290	1.7292
Pan-Mamba	32.3641	6.5956	0.8809	0.9423	2.5166	2.1535
DASFusion	34.6808	5.6421	0.9831	0.9601	2.2612	1.6252
MSPF-Net(Ours)	35.0728	4.6521	0.9624	0.8082	2.0800	1.5096

TABLE II

Quantitative Comparison of Different Methods on the WorldView-3 Dataset Under the RR Protocol

Methods	PSNR $\uparrow$	RMSE $\downarrow$	UIQI $\uparrow$	$Q^{2n} \uparrow$	SAM $\downarrow$	ERGAS $\downarrow$
DRPNN	25.6762	19.5398	0.9224	0.8152	7.9339	6.1439
MSDCNN	25.2981	19.8476	0.9171	0.8033	7.7049	6.2930
PanColorGAN	25.3901	13.9954	0.9149	0.8302	8.3228	5.6929
TFNet	28.2917	12.9021	0.9047	0.8921	5.3725	3.7714
PCDRN	27.8324	13.2502	0.9460	0.8681	5.3713	4.0870
AWFLN	31.7700	11.1031	0.9712	0.9352	4.8546	2.7522
LRTCFPan	28.6875	13.1114	0.9549	0.8987	5.3327	3.6266
WFANet	32.1136	6.0976	0.9352	0.7248	4.5736	2.6858
Pan-Mamba	31.8937	8.0232	0.8582	0.9271	4.8042	3.1505
DASFusion	33.0621	8.7486	0.9795	0.9483	3.9330	2.3133
MSPF-Net(Ours)	32.9968	5.9131	0.9455	0.7628	4.0190	1.1750

TABLE III

Quantitative Comparison of Different Methods on the Pléiades Dataset Under the RR Protocol

Methods	PSNR $\uparrow$	RMSE $\downarrow$	UIQI $\uparrow$	$Q^{2n} \uparrow$	SAM $\downarrow$	ERGAS $\downarrow$
DRPNN	30.1216	11.1515	0.9680	0.8636	2.6809	2.9884
MSDCNN	29.2428	12.2007	0.9629	0.8638	2.7117	3.2713
PanColorGAN	23.1587	13.1564	0.9320	0.8031	7.8692	8.4645
TFNet	32.8707	8.5569	0.9805	0.9452	2.9289	2.0643
PCDRN	31.5914	9.7448	0.9767	0.9004	2.3488	2.6322
AWFLN	35.0748	5.7632	0.9883	0.9652	2.2029	1.5431
LRTCFPan	30.1723	11.7337	0.9697	0.9195	2.8051	2.7336
WFANet	36.5915	8.1019	0.9703	0.8131	2.1990	1.3286
Pan-Mamba	35.4308	4.6590	0.9618	0.9616	2.3408	1.6319
DASFusion	36.8194	5.0668	0.9925	0.9715	2.1060	1.2504
MSPF-Net(Ours)	37.6945	3.3496	0.9751	0.8410	1.9857	1.1464

As shown in Table I, on the IKONOS dataset, MSPF-Net achieves the highest PSNR of 35.0728 dB and the lowest RMSE of 4.6521. Compared with DASFusion, MSPF-Net improves the PSNR by 0.3920 dB and reduces the RMSE by 0.9900. It also obtains the lowest SAM and ERGAS values of 2.0800 and 1.5096, respectively, indicating more accurate reconstruction and lower spectral distortion. Meanwhile, DASFusion achieves higher UIQI and Q^{2n} values of 0.9831 and 0.9601, respectively, compared with 0.9624 and 0.8082 for MSPF-Net, suggesting stronger structural and overall quality consistency according to these metrics. Nevertheless, MSPF-Net outperforms DASFusion, WFANet, and AWFLN in PSNR, RMSE, SAM, and ERGAS, demonstrating its strong pixel-level reconstruction accuracy and spectral fidelity on the IKONOS dataset.

For the WV3 dataset, as reported in Table II, the results reveal a more evident trade-off among the evaluation metrics. DASFusion achieves the highest PSNR of 33.0621 dB, which is only 0.0653 dB higher than the 32.9968 dB obtained by MSPF-Net. However, MSPF-Net achieves the lowest RMSE of 5.9131, reducing the value by 2.8355 compared with DASFusion. It also obtains the lowest ERGAS of 1.1750, which is 1.1383 lower than that of DASFusion. In terms of SAM, DASFusion achieves a slightly lower value of 3.9330, com-

pared with 4.0190 for MSPF-Net. Moreover, DASFusion obtains substantially higher UIQI and Q^{2n} values of 0.9795 and 0.9483, respectively, whereas MSPF-Net achieves 0.9455 and 0.7628. These results suggest that DASFusion has advantages in structural consistency and achieves slightly better PSNR and SAM on WV3, while MSPF-Net exhibits more pronounced advantages in RMSE and ERGAS. In particular, despite the negligible PSNR difference between the two methods, MSPF-Net produces substantially lower reconstruction error and spectral distortion, demonstrating its effectiveness in accurate image reconstruction and spectral fidelity preservation.

On the Pléiades dataset, as shown in Table III, MSPF-Net again demonstrates clear advantages in reconstruction accuracy. It achieves the highest PSNR of 37.6945 dB and the lowest RMSE of 3.3496, improving the PSNR by 0.8751 dB and reducing the RMSE by 1.7172 compared with DASFusion. MSPF-Net also obtains the lowest SAM and ERGAS values of 1.9857 and 1.1464, respectively, corresponding to reductions of 0.1203 and 0.1040 compared with DASFusion. In contrast, DASFusion achieves the highest UIQI and Q^{2n} values of 0.9925 and 0.9715, indicating stronger structural consistency and overall quality according to these metrics. Nevertheless, the superior performance of MSPF-Net in PSNR, RMSE, SAM, and ERGAS demonstrates its ability to achieve more accurate

pixel-level reconstruction while effectively preserving multi-spectral information on the Pléiades dataset.

Overall, the results across the three datasets demonstrate that MSPF-Net maintains stable reconstruction performance under different sensor conditions. DASFusion consistently performs better in UIQI and Q^{2n} and shows slight advantages in PSNR and SAM on WV3, whereas MSPF-Net maintains clear advantages in RMSE and ERGAS and achieves higher PSNR and lower SAM on IKONOS and Pléiades. These results further indicate that different evaluation metrics emphasize different aspects of fusion quality. Overall, MSPF-Net effectively reduces reconstruction errors and spectral distortion while maintaining stable performance across different sensors, demonstrating the

effectiveness and cross-dataset robustness of the proposed multi-scale spatial-spectral feature fusion strategy.

2) Full-Resolution (FR) Quantitative Analysis

To further evaluate the generalization capability of MSPF-Net under real full-resolution conditions, FR experiments are conducted on the Pléiades and WorldView-3 (WV3) datasets, with the quantitative results reported in Table IV. Since no strictly corresponding high-resolution reference images are available in the FR scenario, the no-reference quality metrics D_λ , D_S , and QNR are adopted, where lower D_λ and D_S and higher QNR indicate better fusion quality.

TABLE IV
Quantitative Comparison of Different Methods Under the FR Protocol on the Pléiades(left) and WorldView-3(right) Datasets

Methods	$D_\lambda \downarrow$	$D_S \downarrow$	QNR $\uparrow$	$D_\lambda \downarrow$	$D_S \downarrow$	QNR $\uparrow$
DRPNN	0.0533	0.0363	0.9123	0.0887	0.0869	0.8341
MSDCNN	0.0434	0.0462	0.9123	0.0872	0.0944	0.8287
PanColorGAN	0.0434	0.0618	0.8975	0.1126	0.1138	0.7892
TFNet	0.0318	0.0310	0.9381	0.0163	0.0464	0.9206
PCDRN	0.0375	0.0382	0.9256	0.0602	0.0650	0.8791
AWFLN	0.0243	0.0233	0.9469	0.0196	0.0470	0.9330
LRTCFPan	0.0111	0.0426	0.9468	0.0363	0.0332	0.9319
WFANet	0.0079	0.0425	0.9499	0.0152	0.0252	0.9600
Pan-Mamba	0.0245	0.0254	0.9507	0.0419	0.0565	0.9045
DASFusion	0.0160	0.0331	0.9513	0.0189	0.0162	0.9651
MSPF-Net(Ours)	0.0264	0.0658	0.9094	0.0291	0.0562	0.9162

As shown in Table IV, on the Pléiades dataset, MSPF-Net achieves $D_\lambda = 0.0264$, $D_S = 0.0658$, and $QNR = 0.9094$. These results indicate that MSPF-Net maintains a certain level of spectral consistency under real full-resolution conditions, as reflected by its relatively moderate D_λ . However, its D_S is relatively high compared with the competing methods, indicating that spatial consistency remains a major challenge in the FR setting. In comparison, WFANet achieves the lowest D_λ of 0.0079, AWFLN obtains the lowest D_S of 0.0233, and DASFusion achieves the highest QNR of 0.9513. The considerable differences among these methods suggest that accurately balancing spectral and spatial fidelity remains challenging when applying pansharpening networks to real full-resolution observations.

On the WV3 dataset, MSPF-Net obtains $D_\lambda = 0.0291$, $D_S = 0.0562$, and $QNR = 0.9162$. Compared with Pléiades, MSPF-Net exhibits a similar level of spectral distortion but still shows relatively high spatial distortion. In particular, WFANet achieves the lowest D_λ of 0.0152, whereas DASFusion obtains the lowest D_S of 0.0162 and the highest QNR of 0.9651. Compared with DASFusion, MSPF-Net exhibits higher D_λ , D_S , and lower QNR, indicating that its performance under real WV3 observations still lags behind methods specifically optimized for stronger spatial-spectral consistency. Nevertheless, the relatively moderate D_λ suggests that the spectral infor-

mation inherited from the multispectral observations is not severely degraded, while the larger D_S indicates that the main limitation is associated with spatial consistency.

The discrepancy between the RR and FR results mainly arises from the intrinsic differences in data distribution and supervision conditions. Under the RR protocol, the input observations are synthetically degraded according to the Wald protocol and accompanied by corresponding high-resolution reference images. The network can therefore learn the relationship between spatial detail reconstruction and spectral preservation under explicit supervision. In contrast, FR experiments directly employ the original full-resolution PAN and MS observations acquired by real sensors without strictly corresponding reference images. Sensor response differences, spatial-spectral inconsistencies, imaging noise, and complex scene structures introduce additional uncertainties that are not fully represented by the RR training process. Consequently, the spatial enhancement strategies learned under the RR setting cannot be completely transferred to real FR observations, resulting in increased spatial distortion in the FR evaluation.

Overall, the FR results reveal that MSPF-Net can preserve the spectral characteristics of the input multispectral observations to a certain extent, but its spatial consistency under real full-resolution conditions remains insufficient compared with several advanced methods. In particular, the stronger FR performance of WFANet and DASFusion demonstrates the importance of accurately modeling the relationship between PAN-

derived spatial details and multispectral features under real sensor conditions. These observations complement the RR evaluation by exposing the domain gap between synthetically supervised reconstruction and real full-resolution fusion, and suggest that future improvements should focus on adaptive modeling of sensor-dependent PAN high-frequency structures and stronger spatial-spectral consistency constraints to improve the robustness and generalization capability of MSPF-Net in practical applications.

3) Cross-Model Evaluation with Publicly Available Pretrained Weights

The preceding experiments evaluate the representative state-of-the-art (SOTA) methods under a unified training protocol, where all competing models are retrained on the same datasets and evaluated under consistent experimental settings. Although such comparisons provide a fair assessment of the effectiveness of different network architectures under the same training conditions, they do not fully reflect the generalization capability of these models when applied to unseen data without additional training. In practical remote sensing applications, pretrained models are often directly deployed on new datasets without retraining or parameter fine-tuning. Therefore, it is also important to investigate whether different pansharpening models can maintain their reconstruction capability when transferred to a target dataset under such a setting.

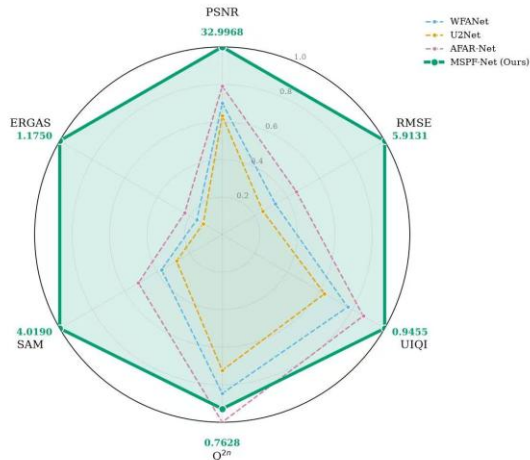


Fig. 5. Radar chart of quantitative performance comparison of different pansharpening models using pretrained weights on the WV3 dataset.

To this end, we further conduct a cross-model evaluation using publicly available pretrained weights. Specifically, WFANet, U2Net, and AFAR-Net are selected as representative competing methods for which pretrained models are publicly available. Unlike the retraining-based comparisons described above, the original pretrained weights of these methods are directly used for inference on the WorldView-3 (WV3) dataset, without any additional training or parameter fine-tuning. All models are evaluated under the same reduced-resolution (RR) protocol, and the corresponding quantitative metrics are calculated using the same reference images. This experiment pro-

vides an additional assessment of the cross-dataset generalization capability of different pansharpening models.

To intuitively demonstrate the overall performance gap among different models, Fig. 5 plots the radar chart of the quantitative results, where all metrics are normalized to relative scores with the optimal value set to 1.0. For metrics where smaller values indicate better performance (i.e., RMSE, SAM, ERGAS), the scores are inversely converted so that the outermost boundary of the radar chart uniformly corresponds to the best performance.

TABLE V
Quantitative comparison of models using publicly available pretrained weights on the WV3 dataset.

Methods	PSNR $\uparrow$	RMSE $\downarrow$	UIQI $\uparrow$	$Q^{2n}\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$
WFANet	23.1582	18.1026	0.7299	0.6957	10.7524	7.5036
U2Net	20.9242	23.6278	0.5956	0.5945	14.2803	10.0662
AFAR-Net	26.0950	12.9448	0.8207	0.8216	7.7910	5.0829
MSPF-Net(Ours)	32.9968	5.9131	0.9455	0.7628	4.0190	1.1750

The quantitative results are reported in Table V. It can be observed that directly applying the publicly available pretrained models to the WV3 dataset leads to considerable performance differences among the competing methods. WFANet, U2Net, and AFAR-Net achieve PSNR values of 23.1582, 20.9242, and 26.0950 dB, respectively, whereas MSPF-Net achieves a substantially higher PSNR of 32.9968 dB. Meanwhile, MSPF-Net obtains the lowest RMSE, SAM, and ERGAS values of 5.9131, 4.0190, and 1.1750, respectively, together with the highest UIQI of 0.9455. These results indicate that MSPF-Net can effectively reduce reconstruction errors while maintaining high structural and spectral consistency when directly applied to WV3 data. In particular, the lower SAM and ERGAS values demonstrate its stronger capability to preserve spectral information while reconstructing high-resolution spatial details.

For the Q^{2n} metric, AFAR-Net achieves the highest value of 0.8216, while MSPF-Net obtains a value of 0.7628. Although MSPF-Net does not achieve the best performance on Q^{2n} , it consistently outperforms the three pretrained competing models in terms of PSNR, RMSE, UIQI, SAM, and ERGAS. This observation suggests that different quality measures may emphasize different aspects of spatial and spectral fidelity. As visually shown in the radar chart, the performance envelope of MSPF-Net is significantly closer to the outermost optimal boundary than the other comparison models, which intuitively verifies its comprehensive advantages in pansharpening quality. Overall, the results demonstrate that MSPF-Net maintains a strong reconstruction capability when directly applied to WV3 without target-domain retraining or parameter fine-tuning, indicating its favorable cross-dataset generalization ability.

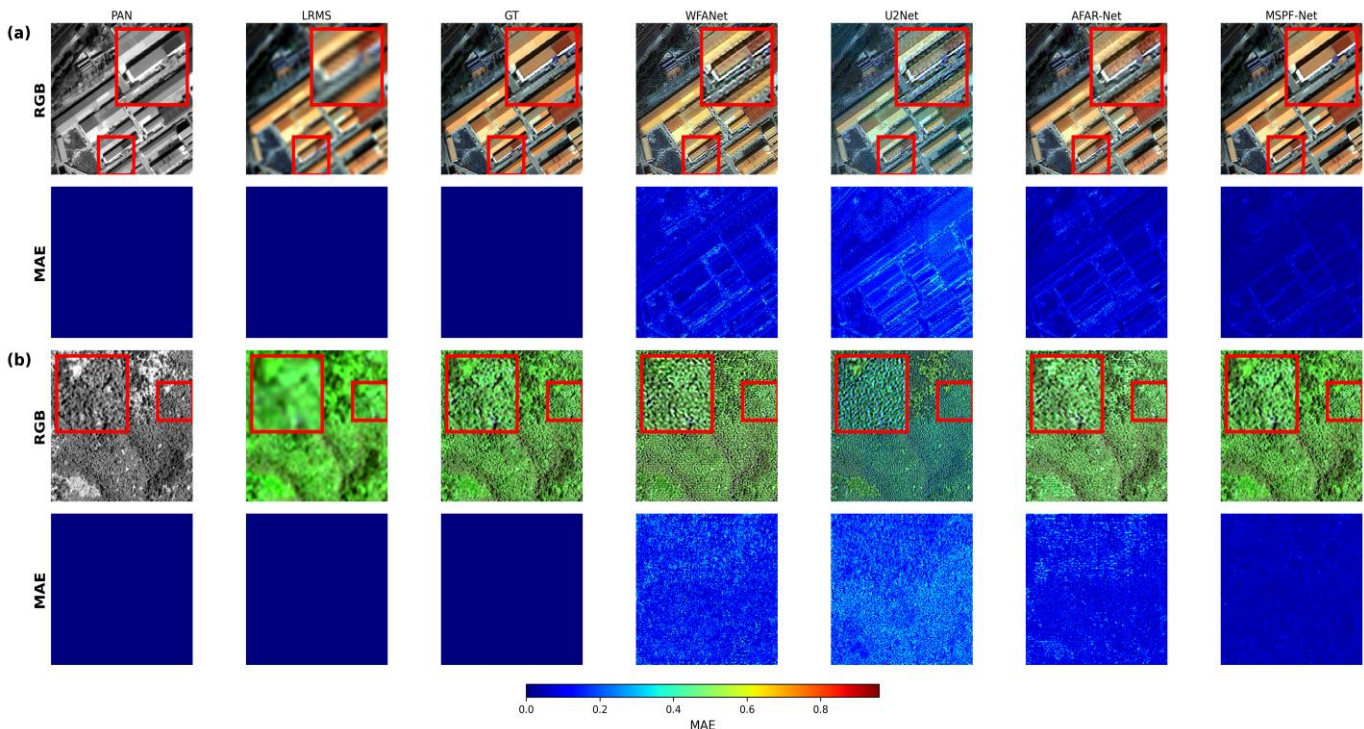


Fig. 6. Visual comparisons of models using pretrained weights on the WV3 dataset, together with the corresponding MAE maps and enlarged local regions. (a) Urban area. (b) Grassland area.

To further provide a visual assessment, Fig. 6 presents representative qualitative comparisons on two WV3 scenes, including an urban area and a grassland area. For each scene, the PAN image, LRMS image, ground-truth (GT) image, and the fusion results obtained by WFANet, U2Net, AFAR-Net, and MSPF-Net are presented, together with the corresponding mean absolute error (MAE) maps and enlarged local regions. The urban scene contains distinct building boundaries, roof structures, and high-frequency textures, making it suitable for evaluating the spatial detail reconstruction capability of different models. As shown in Fig. 6(a), WFANet and U2Net exhibit noticeable structural distortions and detail loss in some building regions, while U2Net produces more pronounced artifacts in the high-frequency areas. AFAR-Net reconstructs relatively sharp structural details but still exhibits certain discrepancies from the GT in some local regions. In comparison, MSPF-Net produces spatial structures and textures that are visually closer to the GT, particularly around building boundaries, roof structures, and fine-scale details.

For the grassland scene shown in Fig. 6(b), the highly heterogeneous vegetation textures and abundant fine-scale spatial information provide a more challenging case for evaluating both spectral preservation and spatial detail reconstruction. WFANet and U2Net exhibit noticeable deviations in local vegetation textures, whereas AFAR-Net provides relatively improved structural reconstruction but still shows discrepancies in several regions. In contrast, MSPF-Net produces a reconstruction with a more consistent overall appearance and finer local textures, which are visually closer to the GT.

The corresponding MAE maps further support these obser-

varations. WFANet and U2Net exhibit relatively strong error responses around building boundaries, texture-rich regions, and vegetation areas, whereas AFAR-Net reduces the overall error to some extent. MSPF-Net presents a more concentrated low-error distribution in both representative scenes, with fewer high-error regions, particularly around complex structures and fine-scale textures. These qualitative observations are consistent with the quantitative results in Table V, demonstrating that MSPF-Net can achieve high-quality pansharpener results under direct inference with its pretrained model and exhibits strong generalization capability on the WV3 dataset.

C. Ablation studies

To validate the effectiveness of each proposed component and investigate their complementary contributions to feature learning, an incremental ablation study is conducted on the Pléiades dataset. The baseline model without any additional proposed modules is first established as Baseline. Subsequently, the proposed modules are progressively incorporated, including the Multi-scale Pyramid Feature Extraction (MPFE) module, PAN High-frequency Enhancement (PHE) module, and Spectral-Spatial Attention Fusion (SSAF) module. Accordingly, four network variants are constructed, namely Baseline, Baseline + MPFE, Baseline + MPFE + PHE, and Baseline + MPFE + PHE + SSAF.

All ablation experiments are performed under the same data split, training strategy, and evaluation protocols to ensure a fair comparison. The quantitative results under the reduced-resolution (RR) and full-resolution (FR) protocols are reported in Table VI and Table VII, respectively.

Table VI
Ablation results of different module configurations on the Pléiades dataset under the RR protocol.

model	PSNR↑(dB)	RMSE↓	UIQI↑	$Q^{2n} \uparrow$	SAM↓(°)	ERGAS↓	SCC↑
Baseline	37.3667±1.0768	3.4774±0.3836	0.974104±0.02824 8	0.836092±0.03726 4	2.0378±0.3631	1.1893±0.2125	0.984749±0.00306 9
Baseline+MPFE	37.4317±1.0767	3.4514±0.3800	0.974267±0.02834 6	0.838754±0.03644 1	2.0303±0.3625	1.1806±0.2108	0.984934±0.00304 4
Base-line+MPFE+PHE	37.6450±1.0977	3.3685±0.3750	0.975022±0.02819 1	0.840218±0.03710 6	1.9925±0.3540	1.1527±0.2036	0.985758±0.00282 1
Base-line+MPFE+PHE+SSAF	37.6945±1.1045	3.3496±0.3747	0.975168±0.02806 7	0.841086±0.03678 2	1.9857±0.3538	1.1464±0.2046	0.985887±0.00277 4

Table VII
Ablation results of different module configurations on the Pléiades dataset under the FR protocol.

model	$D_\lambda \downarrow$	$D_s \downarrow$	QNR↑
Baseline	0.026125±0.007764	0.072290±0.021494	0.903398±0.018793
Baseline+MPFE	0.023277±0.007480	0.067539±0.021413	0.910674±0.018298
Baseline+MPFE+PHE	0.026616±0.007916	0.084306±0.024086	0.891234±0.021072
Baseline+MPFE+PHE+SSAF	0.026428±0.007412	0.065824±0.022137	0.909417±0.019552

1) Progressive Ablation Analysis under the RR Protocol

Table VI reports the quantitative results of four progressively configured network variants on the Pléiades dataset under the RR protocol. Overall, the gradual incorporation of MPFE, PHE, and SSAF leads to consistent performance improvements across multiple evaluation metrics, demonstrating the effectiveness and complementary contributions of the proposed components.

Compared with the Baseline, introducing the Multi-scale Pyramid Feature Extraction (MPFE) module improves the PSNR from 37.3667 dB to 37.4317 dB, while reducing the RMSE from 3.4774 to 3.4514. Meanwhile, Q^{2n} , SAM, ERGAS, and SCC are improved from 0.836092, 2.0378°, 1.1893, and 0.984749 to 0.838754, 2.0303°, 1.1806, and 0.984934, respectively. Although the numerical improvement is relatively limited, the consistent enhancement across all evaluation metrics indicates that MPFE effectively strengthens the representation of spatial structures by introducing multi-scale contextual information. Such hierarchical feature modeling provides richer structural priors for subsequent feature fusion, particularly in complex spatial scenarios.

With the further incorporation of the PAN High-frequency Enhancement (PHE) module, the reconstruction performance is further improved. Specifically, the PSNR increases from 37.4317 dB to 37.6450 dB, corresponding to a gain of 0.2133 dB, while the RMSE decreases from 3.4514 to 3.3685. In addition, UIQI, Q^{2n} , and SCC increase to 0.975022, 0.840218, and 0.985758, respectively, whereas SAM and ERGAS are reduced to 1.9925° and 1.1527. Compared with MPFE, PHE provides a

more significant performance improvement, demonstrating that explicitly enhancing high-frequency spatial components from PAN features can effectively strengthen the recovery of edges, textures, and local spatial details.

After integrating the Spectral-Spatial Attention Fusion (SSAF) module, the complete model achieves further improvements across all RR evaluation metrics. The PSNR is increased from 37.6450 dB to 37.6945 dB, and the RMSE is reduced from 3.3685 to 3.3496. Meanwhile, Q^{2n} and SCC are further improved to 0.841086 and 0.985887, while SAM and ERGAS are reduced to 1.9857° and 1.1464, respectively. The final MSPF-Net achieves the best performance among all four configurations, indicating that SSAF effectively enhances the interaction between spatial and spectral-related features after multi-scale representation and high-frequency enhancement.

Overall, the progressive improvements from MPFE, PHE, to SSAF demonstrate their distinct yet complementary roles within MSPF-Net. MPFE mainly establishes hierarchical multi-scale spatial representations, PHE enhances informative spatial details from PAN features, and SSAF further optimizes the interaction between spatial and spectral information through adaptive feature refinement. The collaborative effect of these modules enables the network to achieve more accurate spatial reconstruction while maintaining spectral consistency under the RR protocol.

2) Module Collaboration Analysis under the FR Protocol

Table VII further evaluates the four configurations under the FR protocol. Compared with the RR results, the FR evaluation reveals a more complex interaction among the proposed modules,

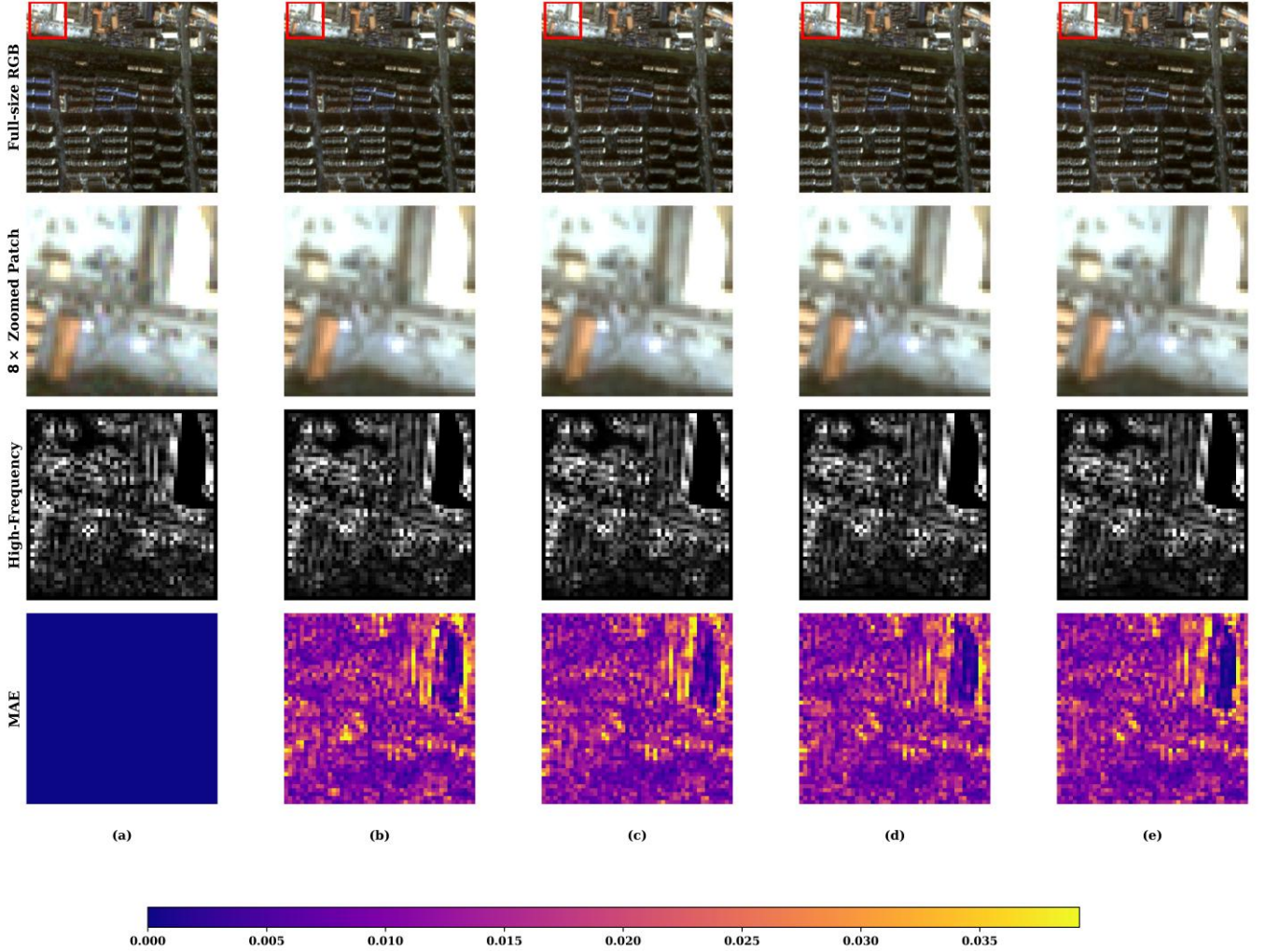


Fig. 7. Visual comparison of ablation study results on the Pléiades dataset under the RR protocol. From top to bottom: full-size fused RGB image, $8\times$ zoomed patch, high-frequency component map, and mean-absolute-error (MAE) map. (a) Ground-truth reference; (b) Baseline; (c) Baseline + MPFE; (d) Baseline + MPFE + PHE; (e) Baseline + MPFE + PHE + SSAF. The color bar denotes the MAE magnitude.

articularly under the condition without high-resolution reference supervision.

Compared with the Baseline, introducing MPFE consistently improves all FR metrics. Specifically, D_λ decreases from 0.026125 to 0.023277, D_s decreases from 0.072290 to 0.067539, and QNR increases from 0.903398 to 0.910674. These improvements indicate that multi-scale spatial representation is beneficial not only for reference-based reconstruction but also for real full-resolution fusion, where richer spatial-context information helps maintain better spatial-spectral consistency.

However, when PHE is further introduced into the MPFE-based model, the FR performance shows a different trend. The values of D_λ and D_s increase from 0.023277 to 0.026616 and from 0.067539 to 0.084306, respectively, while QNR decreases from 0.910674 to 0.891234. This phenomenon indicates that explicitly enhancing PAN high-frequency responses may introduce stronger spatial priors, which are beneficial for detail reconstruction under the RR protocol but may cause spatial inconsistency when evaluated in the unconstrained FR scenario. Therefore, high-frequency enhancement should be properly regulated to avoid excessive spatial information injection.

After incorporating SSAF, the FR performance is effectively recovered. Specifically, D_λ decreases from 0.026616 to 0.026428, while D_s is significantly reduced from 0.084306 to 0.065824, corresponding to a relative reduction of approximately 21.93%. Meanwhile, QNR increases from 0.891234 to 0.909417. Notably, the final model achieves a lower D_s than the MPFE configuration (0.067539), demonstrating that SSAF can effectively suppress the spatial inconsistency introduced by high-frequency enhancement and achieve a better balance between spatial detail preservation and spectral fidelity.

It is worth noting that the complete model does not achieve the optimal D_λ or QNR among all configurations. Rather than indicating a limitation of SSAF, this result reflects the inherent trade-off between spatial enhancement and spectral consistency in full-resolution pansharpening. The primary contribution of SSAF is to adaptively regulate the interaction between enhanced spatial features and spectral-related representations, thereby preventing excessive spatial responses while maintaining effective detail preservation.

Overall, the FR ablation results provide complementary evi-

dence to the RR evaluation. MPFE mainly improves the robustness of spatial-context modeling, PHE enhances the representation of high-frequency spatial structures, while SSAF plays a crucial role in coordinating the enhanced spatial information with spectral features. The progressive integration of these modules enables MSPF-Net to achieve a more balanced spatial-spectral reconstruction capability under both reference-based and real full-resolution scenarios.

3) Qualitative Visual Ablation Analysis

To provide intuitive visual evidence for the quantitative results reported in Table VI, qualitative ablation comparisons are conducted on the Pléiades dataset under the RR protocol, as shown in Fig. 7. The first row shows full-size fused images, with red rectangles denoting the regions selected for local zoom-in inspection. The second row presents $8\times$ magnified patches to reveal fine spatial structures such as edges and textures. The third row visualizes high-frequency maps that reflect the edge and texture responses of reconstructed outputs. The fourth row depicts local mean absolute error (MAE) maps, in which brighter pixels represent larger reconstruction deviations relative to the ground-truth reference.

As can be observed, the Baseline suffers from evident texture blurring, and prominent reconstruction errors appear in complex-structure regions in the MAE maps. After embedding the MPFE module, the reconstructed results preserve richer contextual information and alleviate partial blurring artifacts. Further integration of the PHE module sharpens edge contours and texture details, which aligns with the quantitative improvements in Table VI. With the additional SSAF module, spatial-spectral feature interaction is further optimized, visual distortions are greatly suppressed, and the overall error magnitude in MAE maps is reduced.

Collectively, visual observations from RGB outputs, high-frequency maps and error maps are highly consistent with the numerical trends of RR-based quantitative metrics. These qualitative comparisons further verify the complementary effects of MPFE, PHE, and SSAF. Progressive integration of these modules jointly boosts the spatial-spectral reconstruction performance for pansharpening.

4) Spectral Reconstruction and Convergence Analysis

To further investigate the spectral reconstruction capability of different module configurations, Fig. 8 presents the per-band PSNR results on the Pléiades dataset under the RR protocol. The results demonstrate that the progressive incorporation of MPFE, PHE, and SSAF consistently improves the reconstruction quality across all four spectral bands. Specifically, the complete model achieves the highest PSNR values in every spectral channel, indicating that the proposed modules contribute to a balanced enhancement rather than improving only specific bands.

This behavior suggests that the proposed spatial-spectral feature learning strategy effectively preserves spectral consistency while enhancing spatial details. By jointly modeling multi-scale spatial structures, high-frequency spatial information, and adaptive feature interactions, MSPF-Net achieves more uniform spectral reconstruction performance across different bands.

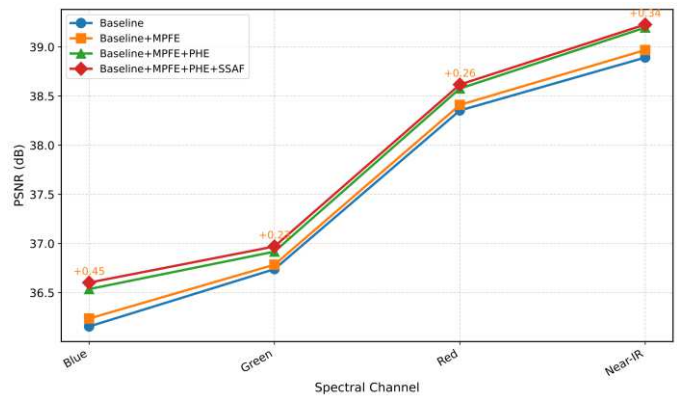


Fig. 8. Per-band PSNR comparison of different module configurations on the Pléiades dataset under the RR protocol.

Fig. 9 further illustrates the training and validation loss curves of the proposed model. The loss decreases rapidly during the early training stage and gradually converges as optimization proceeds. The training and validation curves exhibit consistent trends without significant oscillations or divergence, demonstrating stable optimization behavior and effective generalization capability.

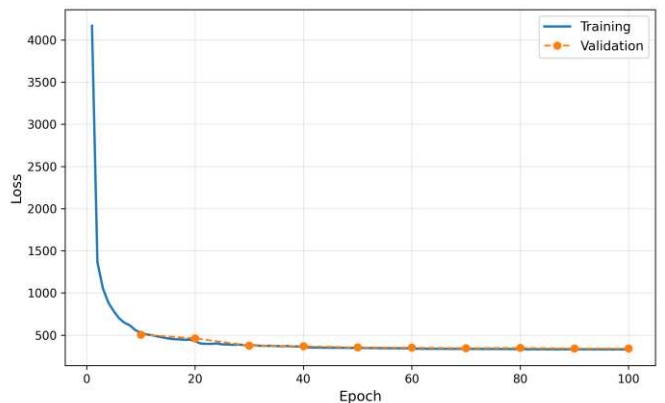


Fig. 9. Training and validation loss curves of the proposed model on the Pléiades dataset.

The stable convergence indicates that the proposed combination of multi-scale feature extraction, high-frequency enhancement, and spectral-spatial adaptive fusion can be effectively optimized under the adopted training strategy.

V. CONCLUSION

This paper proposes a multi-scale spatial-spectral feature fusion network (MSPF-Net) for pansharpening, aiming to address the challenging trade-off between spatial detail enhancement and spectral information preservation. By integrating multi-scale spatial representation, PAN high-frequency enhancement, and spectral-spatial adaptive fusion, the proposed framework establishes a progressive feature learning strategy that effectively exploits complementary information from PAN and multispectral images.

Extensive experiments conducted on the IKONOS, WorldView-3, and Pléiades datasets demonstrate the effectiveness of MSPF-Net. Under the RR protocol, the proposed method achieves consistently superior performance and obtains the best results in terms of PSNR, RMSE, SAM, and ERGAS

across all three datasets, indicating its strong capability in spatial detail reconstruction and spectral fidelity preservation. Furthermore, the FR experiments verify the applicability of MSPF-Net under real full-resolution scenarios, showing its potential for cross-sensor pansharpening.

Although MSPF-Net achieves competitive performance, improving fusion quality under real full-resolution conditions remains challenging due to the absence of explicit reference information and the complex degradation characteristics of practical remote sensing images. Future work will focus on more accurate degradation modeling, cross-domain feature adaptation, and efficient spatial-spectral interaction mechanisms to further enhance the robustness and generalization capability of pansharpening models in complex real-world scenarios.

REFERENCES

- [1] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges," *Information Fusion*, vol. 46, pp. 102–113, 2019.
- [2] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2014.
- [3] T.-M. Tu, P. S. Huang, C.-L. Hung, and C.-P. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 4, pp. 309–312, 2004.
- [4] X. Tian, Y. Chen, C. Yang, and J. Ma, "Variational pansharpening by exploiting cartoon-texture similarities," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.
- [5] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, pp. 594, 2016.
- [6] L.-J. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 3, pp. 279–315, 2022.
- [7] Y. Que, H. Xiong, X. Xia, J. You, and Y. Yang, "Integrating spectral and spatial bilateral pyramid networks for pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 3985–3998, 2024.
- [8] L. Jian, S. Wu, L. Chen, G. Vivone, R. Rayhana, and D. Zhang, "Multi-scale and multi-stream fusion network for pansharpening," *Remote Sensing*, vol. 15, no. 6, pp. 1666, 2023.
- [9] S. Xu, S. Zhong, H. Li, and C. Gong, "Spectral-Spatial Attention-Guided Multi-Resolution Network for Pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 7559–7571, 2025.
- [10] X. Wen, H. Ma, and L. Li, "A three-branch pansharpening network based on spatial and frequency domain interaction," *Remote Sensing*, vol. 17, no. 1, pp. 13, 2024.
- [11] L. Jian, J. Liu, L. Chen, D. Zhang, G. Vivone, and X. Zhou, "Feature interaction and adaptive fusion network with spectral modulation for pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 10458–10470, 2025.
- [12] L. Jian, Y. Chen, G. Vivone, and Y. Zheng, "Cross-modal information aggregation network with feature enhancement for pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 19–33, 2025.
- [13] R. Sustika, A. B. Suksmono, D. Danudirdjo, and K. Wikantika, "Remote sensing image pansharpening using deep internal learning with residual double-attention network," *IEEE Access*, vol. 12, pp. 162285–162298, 2024.
- [14] Z. Li, G. Yuan, and J. Li, "DUCD: Deep Unfolding Convolutional-Dictionary network for pansharpening remote sensing image," *Expert Systems with Applications*, vol. 249, pp. 123589, 2024.
- [15] H. Tao, G. Yuan, Z. Hua, and J. Li, "PDAF: Prompt-driven dynamic adaptive fusion network for pansharpening remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 13533–13546, 2024.
- [16] X. Feng, J. Wang, Z. Zhang, and X. Chang, "Remote sensing image pan-sharpening via Pixel difference enhance," *International Journal of Applied Earth Observation and Geoinformation*, vol. 132, pp. 104045, 2024.
- [17] S. Xu, S. Zhong, and C. Gong, "A Robust Frequency Domain-Generalized Network for Multispectral Pansharpening," *IEEE Journal of Selected Topics in Signal Processing*, 2025.
- [18] Q. Zhou, W. Li, Y. Peng, and X. Zhi, "MSSDIBNet: Multiple spatial-spectral dual-injection balance network for pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 1–18, 2025.
- [19] Y. Yang, M. Li, S. Huang, W. Wan, H. Lu, and W. Tu, "DASFusion: Diffusion-Enhanced Asymmetric Scale-Gated Fusion Network for Pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, 2026.
- [20] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1795–1799, 2017.
- [21] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "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*, vol. 11, no. 3, pp. 978–989, 2018.
- [22] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110–120, 2020.
- [23] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, vol. 55, pp. 1–15, 2020.
- [24] Y. Yang, W. Tu, S. Huang, and H. Lu, "PCDRN: Progressive cascade deep residual network for pansharpening," *Remote Sensing*, vol. 12, no. 4, pp. 676, 2020.
- [25] H. Lu, Y. Yang, S. Huang, X. Chen, B. Chi, A. Liu, and W. Tu, "AWFLN: An adaptive weighted feature learning network for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

- [26] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J. Huang, J. Chanussot, and G. Vivone, "LRTCFFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Transactions on Image Processing*, vol. 32, pp. 1640–1655, 2023.
- [27] J. Huang, R. Huang, J. Xu, S. Peng, Y. Duan, and L.-J. Deng, "Wavelet-assisted multi-frequency attention network for pansharpening." pp. 3662–3670.
- [28] X. He, K. Cao, J. Zhang, K. Yan, Y. Wang, R. Li, C. Xie, D. Hong, and M. Zhou, "Pan-mamba: Effective pan-sharpening with state space model," *Information Fusion*, vol. 115, pp. 102779, 2025.