Progressive High-Frequency Reconstruction for Pan-Sharpening with Implicit Neural Representation

Ge Meng[1], Jingjia Huang[1], Yingying Wang[2], Zhenqi Fu[1], Xinghao Ding[1, 2*], Yue Huang[1, 2]

[1]School of Informatics, Xiamen University, China
[2]Institute of Artificial Intelligence, Xiamen University, China
{mengg, huangjj, wangyingying7, fuzhenqi}@stu.xmu.edu.cn, {dxh, yhuang2010}@xmu.edu.cn

Abstract

Pan-sharpening aims to leverage the high-frequency signal of the panchromatic (PAN) image to enhance the resolution of its corresponding multi-spectral (MS) image. However, deep neural networks (DNNs) tend to prioritize learning the low-frequency components during the training process, which limits the restoration of high-frequency edge details in MS images. To overcome this limitation, we treat pan-sharpening as a coarse-to-fine high-frequency restoration problem and propose a novel method for achieving high-quality restoration of edge information in MS images. Specifically, to effectively obtain fine-grained multi-scale contextual features, we design a Band-limited Multi-scale High-frequency Generator (BMHG) that generates high-frequency signals from the PAN image within different bandwidths. During training, higher-frequency signals are progressively injected into the MS image, and corresponding residual blocks are introduced into the network simultaneously. This design enables gradients to flow from later to earlier blocks smoothly, encouraging intermediate blocks to concentrate on missing details. Furthermore, to address the issue of pixel position misalignment arising from multi-scale features fusion, we propose a Spatial-spectral Implicit Image Function (SIIF) that employs implicit neural representation to effectively represent and fuse spatial and spectral features in the continuous domain. Extensive experiments on different datasets demonstrate that our method outperforms existing approaches in terms of quantitative and visual measurements for high-frequency detail recovery.

Introduction

Compared with single-channel or RGB images, multispectral (MS) images contain richer spectral information. Therefore, MS imaging has been widely used in environmental monitoring, agriculture, mapping services and so on. Nevertheless, the physical constraints of satellites impede the direct capture of high-resolution multispectral (HRMS) images via sensors. As a substitute, they can only obtain high-resolution panchromatic (PAN) images along with their corresponding low-resolution multispectral (LRMS) images. Pan-sharpening is a technique that injects the high-resolution information from the PAN image into its corresponding LRMS image to generate the HRMS image.

*Corresponding Author.

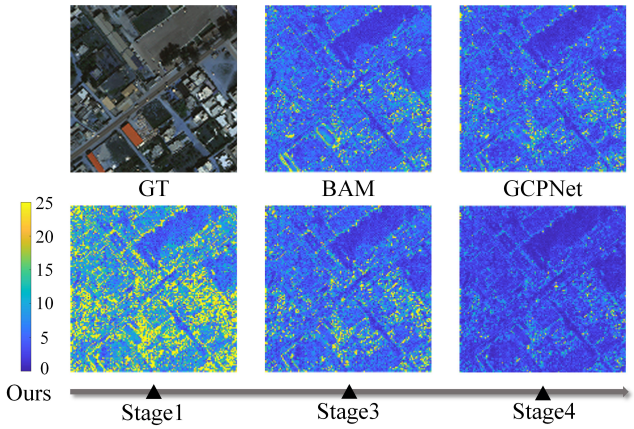


Figure 1: Injecting high-frequency signals into MS image progressively. The residual map between the predicted result and ground truth shows that as the training stage progresses, finer edge details are reconstructed.

The high-quality restoration of high-frequency information in MS images is the core task of pan-sharpening. However, during the training process, deep neural networks (DNNs) often learn to fit target functions from low to high frequencies (Xu et al. 2019; Xu, Zhang, and Xiao 2019), which is referred to as spectral bias (Rahaman et al. 2019). This tendency can restrict the network's ability to capture fine edge details. Training models in a meaningful order may enhance their effectiveness in tackling challenging tasks (Dai et al. 2020; Guo et al. 2018; Guo et al. 2020; Karras et al. 2017; Soviany et al. 2022; Zhou, Wang, and Bilmes 2021). Therefore, we regard pan-sharpening as a coarse-to-fine high-frequency reconstruction problem. Figure 1 displays the fusion results of HRMS images at different training stages. It could be observed that as the training stage progresses, finer edge details are reconstructed.

To effectively obtain multi-scale contextual features, some approaches (Khan et al. 2008; Otazu et al. 2005; Ranchin and Wald 2000; Aiazzi et al. 2006) attempt to decompose the PAN and LRMS images by using multiresolution tools, such as wavelets and Laplacian pyramids. Subsequently, they fuse the generated features to achieve a com-

prehensive representation. However, due to the disparities in high-frequency regions, the results generated by fusing these decomposed components often suffer from aliasing and local dissimilarities. Existing CNN-based multi-scale feature extraction methods (Wang et al. 2023, Fu et al. 2020) represent spatial and spectral features in the discrete domain, which can disrupt the distribution characteristics of the physical signals. Moreover, the feature extraction capability is severely limited by the simple combination of dilated convolution operations with different dilated factors (Fu et al. 2020). Simply repeatedly applies nonlinear filters (such as a sinusoid function) to the network's input, then multiplies the linear functions of these features together, signals at multiple scales can be effectively represented (Fathony et al. 2020; Lindell et al. 2022). Inspired by this, we design a Band-limited Multi-scale High-frequency Generator (BMHG) that generates high-frequency signals from the PAN image within different bandwidths. Specifically, by explicitly controlling the bandwidth of the input signals, different scales of high-frequency signals can be generated that approximate a normal distribution. As the training progresses, higher-frequency signals are generated and progressively injected into the MS image.

In MS images, the different spectra are approximately continuous. However, sensors typically store and represent an MS image as a 3D array of pixels, disrupting the continuity of the physical signals. By using CNN to represent and fuse multi-scale features in the discrete domain (Fu et al. 2020), pixel-level alignment cannot be precisely achieved, leading to the appearance of artifacts and edge misalignment in the fused results. Benefits from its powerful ability to represent signals in the continuous domain (Sitzmann et al. 2020; Tancik et al. 2020), implicit neural representation (INR) has achieved impressive results in the fields of image super-resolution (Chen, Liu, and Wang 2021), scene rendering (Mildenhall et al. 2021; Niemeyer et al. 2020), and 3D reconstruction (Jiang et al. 2020; Park et al. 2019; Saito et al. 2019) in recent years. Additionally, in the realm of image representation, employing periodic functions (Sitzmann et al. 2020), rather than traditional activation functions like ReLU, offers an effective approach for modeling signals with fine details. To represent HRMS images with high fidelity, we approach the problem of fusing spatial and spectral information in pan-sharpening from the perspective of INR. We design the Spatial-spectral Implicit Image Function (SIIF) for representing HRMS images in a continuous manner. In SIIF, a HRMS image is represented as a set of latent codes distributed in 2D spatial dimensions and 1D spectral dimension. Given a 2D spatial coordinate and a 1D spectral coordinate, the decoding function takes the coordinate information and queries the local latent codes around the coordinate as inputs, then predicts the RGB value at the given coordinate as an output.

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

- We propose a novel pan-sharpening method that utilizes a progressive strategy to fuse spatial and spectral features in the continuous domain. This approach effectively resolves the difficulty of accurately fitting high-frequency functions of DNNs during the learning process.
- We design a multi-stage band-limited signal generation network called BMHG. This network generates multi-scale high-frequency signals that closely approximate the distribution characteristics of the physical world by explicitly controlling the bandwidth of the input signals.
- We design an implicit image representation function SIIF that effectively fuses the features of two kinds of modality images in both the continuous 2D spatial domain and continuous 1D spectral domain. The ablation experiments also prove the effectiveness of SIIF.
- Extensive experiments on three satellite datasets demonstrate that our proposed method outperforms existing methods in both quantitative and qualitative metrics.

Methodology

Figure 2 shows the overall architecture of our Progressive Implicit Feature Fusion Network (PIF-Net), which mainly consists of three parts, Band-limited Multi-scale High-frequency Generator (BMHG), Spatial-spectral Implicit Image Function (SIIF), and Progressive High-frequency Injection Module (PHIM). The details will be illustrated below.

Band-limited Multi-scale High-frequency Generator

To ensure accurate recovery of edge information for objects of different sizes in MS images, it is necessary to effectively obtain fine-grained multi-scale contextual features. Subsequently, these multi-scale features need to be stretched to the same resolution to fit convolutional operations. However, during this process, simple bilinear or bicubic interpolations can easily result in artifacts and misalignment. Therefore, it is crucial to eliminate misalignment arising from different scale features. Inspired by (Lindell et al. 2022), we manipulate the bandwidth of input signals explicitly to enable the network to generate high-frequency signals of different scales. Specifically, given the PAN image $P \in R^{H \times W \times 1}$ and LRMS image $M \in R^{\frac{H}{r} \times \frac{W}{r} \times C}$, their edge information is first extracted:

$$\begin{aligned} I_{PAN}^{hp} &= \nabla(P) \\ I_{M\uparrow}^{hp} &= \mathrm{UP}(\nabla(M)) \end{aligned} \tag{1}$$

where $\mathrm{UP}(\cdot)$ represents upsampling the gradient image of M to match the resolution of P, and ∇ represents gradient operation. Then, gradient images of two kinds of modality images can be obtained. Next, we derive the frequency-domain signal of the PAN image using positional encoding:

$$\begin{aligned} \mathcal{P}(x) = (&sin(2^0 \cdot x \cdot \pi), cos(2^0 \cdot x \cdot \pi), \cdots, \\ &sin(2^{L-1} \cdot x \cdot \pi), cos(2^{L-1} \cdot x \cdot \pi)), \\ &L = \log_2 N \end{aligned} \tag{2}$$

Here, x represents the 2D spatial coordinate of a pixel, L represents the high-order harmonic indices of the signal and N represents the spatial resolution of P. We initialize BMHG to have a maximum bandwidth B of 0.5 cycles/pixel, which is the Nyquist limit for the image.

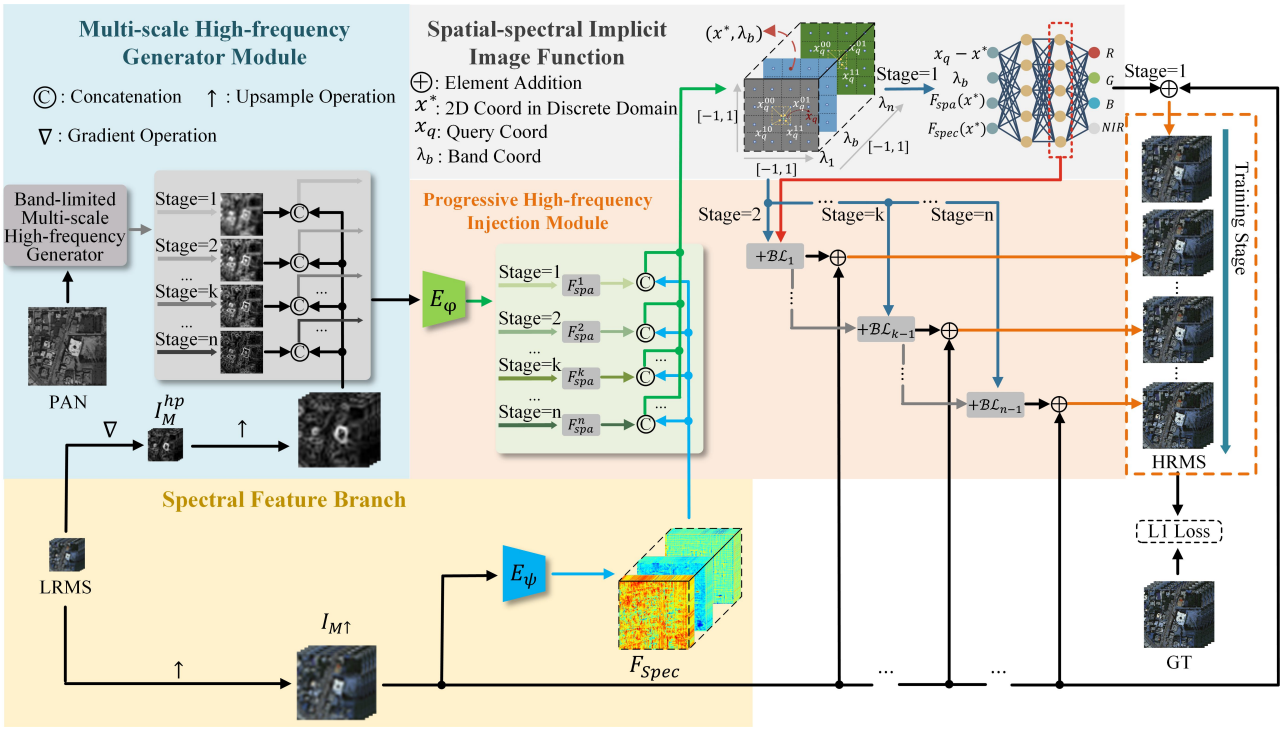
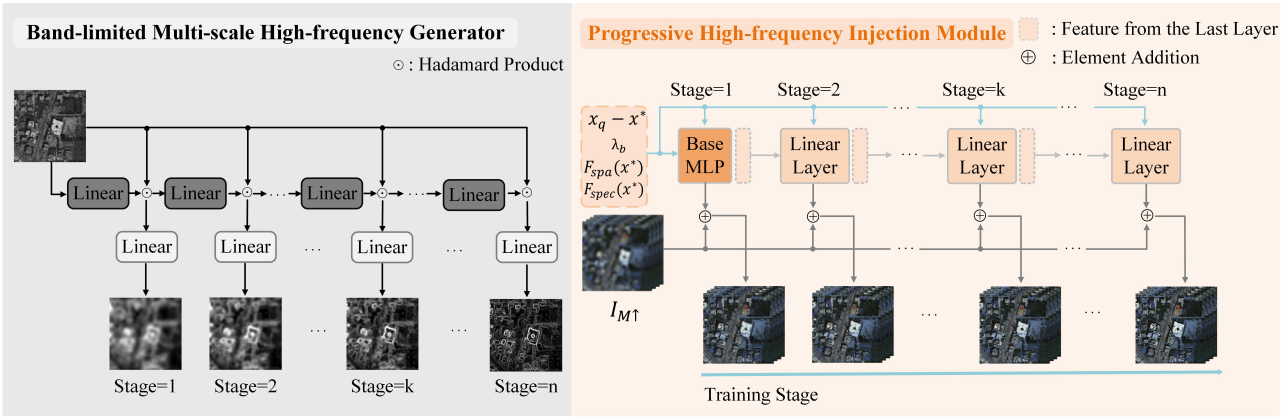


Figure 2: The overall framework of PIF-Net. It consists of three key parts: Multi-scale High-frequency Generator Module, Spatial-spectral Implicit Image Function and Progressive Information Injection Module.



(a) Band-limited Multi-scale High-frequency Generator

(b) Progressive High-frequency Injection Module

Figure 3: Architectures of the proposed two key modules. The sub-figure (a) and (b) depict the BMHG and PHIM respectively.

We design a multi-stage band-limited signal generation network, and initialize input signals for each stage by limiting their frequency range, and use linear layers to achieve signal interaction between different stages during the forward propagation of the network. The structure of BMHG is shown in Figure 3 (a):

$$\begin{aligned} s_i &= \mathcal{P}_{-B_i}^{B_i} \\ h_0 &= s_0 \\ h_i &= s_i \odot (\mathbf{W}_i s_{i-1} + \mathbf{b}_i) \\ y_i &= \mathbf{W}_i^{out} h_i + \mathbf{b}_i^{out}, i = 0, 1, \cdots, n \end{aligned} \tag{3}$$

where s_i represents the input signal of the PAN image in the i-th stage and B_i represents the bandwidth of s_i, $\mathbf{W}_i$ and $\mathbf{b}_i$ represent the weight and bias of the i-th hidden linear layer (dark gray block), while $\mathbf{W}_i^{out}$ and $\mathbf{b}_i^{out}$ represent the weight and bias of the i-th output layer (light gray block), h_i represents the intermediate signal generated in the i-th stage and $\odot$ indicates the Hadamard product.

The input signal from the previous stage first passes through a linear layer, then interacts with the current input signal, and finally the output signal y_i is generated after passing through another linear layer, y_i also can be ex-

pressed as:

$$y_i = \sum_{j=0}^{N_i-1} \boldsymbol{\alpha}_j \cdot \sin\left(\boldsymbol{\phi}_j\right)$$
$$N_{n+1} = \sum_{i=0}^{n} 2^i \cdot d_h^{i+1} \tag{4}$$

This means that each signal from the output layer can be represented as a sum of multiple sine waves with amplitude of $\boldsymbol{\alpha}_j$ and phase of $\boldsymbol{\phi}_j$, where the bandwidth of $\boldsymbol{\phi}_j$ depends on the frequency of the input signals of all previous stages, d_h is the dimension of the signal generated from the hidden layer. The distribution of y_i from each stage is approximately zero-mean Gaussian with variance:

$$var\left(\boldsymbol{\omega}_i\right) \cdot \sum_{m=0}^{n} m \cdot \frac{2^{n-m} d_h^{n+1-m}}{\sum_{i=0}^{n} 2^i d_h^{i+1}} \tag{5}$$

where $\boldsymbol{\omega}_i$ is the frequency of $\boldsymbol{\phi}_i$ and $var(\cdot)$ means the variance. The ideal multi-scale high-frequency information for P can be generated at each stage under the supervision of I_{PAN}^{hp}. Furthermore, the high-frequency signals of different scales from the PAN image, along with the high-frequency signals from the LRMS image, will be jointly fed into the next stage of the network:

$$HP(P, M) = concat(y, I_M^{hp}) \tag{6}$$

Spatial-spectral Implicit Image Function

We apply two encoder networks to extract spatial feature and spectral feature from the PAN image and the MS image respectively:

$$F_{spa} = E_\varphi(HP(P, M))$$
$$F_{spec} = E_\psi(\mathrm{UP}(M)) \tag{7}$$

where E_φ and E_ψ are two encoder networks with parameters φ and ψ respectively, their details are shown in Figure 4. Increasing the number of feature maps beyond a certain level would result in numerical instability during the training procedure (Szegedy et al. 2017; Lim et al. 2017). So we adopt residual scaling with a factor of 0.1. In each residual block, constant scaling layers (Mult block) are placed after the last convolution layers.

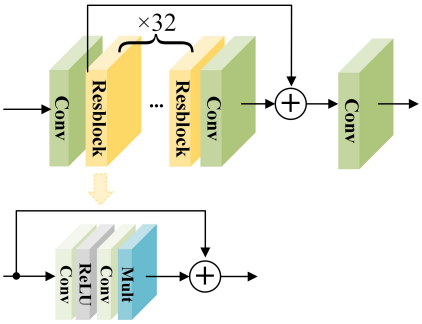


Figure 4: The detail of encoder E_φ and E_ψ.

After obtaining the spatial feature F_{spa} and the spectral feature F_{spec}, we further fuse them in the continuous domain to achieve accurate pixel-level alignment. The spatial-spectral feature in pan-sharpening can be seen as latent codes distributed in the continuous 3D space, with each feature vector assigned a 2D spatial coordinate and a 1D spectral coordinate. The latent code at the coordinate x^* in the discrete feature map, which is closest to the query coordinate x_q in the continuous feature map, can be represented as:

$$I(x^*) = f_\theta(F_{spa}(x^*),\, F_{spec}(x^*),\, x_q - x^*,\, \lambda_b) \tag{8}$$

where $x_q - x^*$ represents the offset between x^* and x_q, $F_{spa}(x^*)$ and $F_{spec}(x^*)$ represent the spatial and spectral feature value at x^* respectively, λ_b represents the band index to which the feature map belongs, which is normalized to [-1, 1], and f_θ represents an MLP network with parameters θ. Figure 2 illustrates the approach for calculating the pixel value at query coordinate x_q. Through the neighboring nodes of the query coordinate x_q in each feature map, we can derive its pixel value:

$$v_q = \sum_{i \in \{00,01,10,11\}} \frac{S_q^i}{S} \cdot I(x_q^i) \tag{9}$$

where $i \in \{00, 01, 10, 11\}$ represents the four neighboring nodes of x_q (top-left, top-right, bottom-left, and bottom-right), S_q^i represents the rectangular area formed by x_q and x_q^i, and $S = \sum_{i \in \{00,01,10,11\}} S_q^i$.

Progressive High-frequency Injection Module

DNNs often fit target functions from low to high frequencies during the training process. In another word, at the early stage of training, the low-frequencies are fitted and as iteration steps of training increase, the high-frequencies are fitted (Xu, Zhang, and Luo 2022). As a result, the learning of high-frequency information would be limited. In order to inject finer high-frequency details from the PAN image into the MS image more effectively, we design a Progressive High-frequency Injection Module (PHIM). We divide the training process into several stages. The training starts with a single-scale coarse high-frequency signal, and then progressively adds higher-frequency signals into new stages. As shown in Figure 3 (b), for the high-frequency signal y_i generated from the i-th output layer of the BMHG, it passes through a block containing two linear layers together with the feature of the previous stage of the network:

$$c_i = \mathcal{BL}_i(c_{i-1},\, F_{spa}^i), i = 1, \cdots, n$$
$$c_0 = \imath_b \tag{10}$$

where $\imath_b$ is the feature derived from the last hidden layer of the base block. The additional block $\mathcal{BL}_i$ outputs a finer result based on the latent features c_{i-1} obtained from the last mapping layer of the previous block. $\mathcal{BL}_i$ has mutual benefits for all scales, and it enables gradients obtained from latter blocks to smoothly flow back to earlier blocks and encourages intermediate blocks to concentrate on the missing details.

Loss Function

In this paper, we use L1 loss to optimize the fusion results in each stage at the pixel level:

$$\mathcal{L}1(\widehat{X}, X) = \frac{1}{n}\sum_{i}^{n} |\widehat{X}(i) - X(i)| \tag{11}$$

where n is the total number of sampled pixels, $\widehat{X}(i)$ represents the predicted value for pixel i, and $X(i)$ represents its corresponding ground truth. We use L2 loss to optimize the BMHG:

$$\mathcal{L}2(y_k, I_{PAN}^{hp}) = ||y_k - I_{PAN}^{hp}||_2^2 \tag{12}$$

where y_k is the high-frequency signal generated by the k-th output layer.

Experiments

Baseline Methods

To demonstrate the effectiveness of our method, we compare the performance of our method with several state-of-the-art methods. We select five traditional pan-sharpening methods, including SFIM (Liu 2002), Brovey (Gillespie, Kahle, and Walker 1987), GS (Laben and Brower 2000), IHS (Carper et al. 1990), and GFPCA (Liao et al. 2015), and five deep learning-based pan-sharpening methods, including BAM (Jin et al. 2021), SFIIN (Zhou et al. 2022), GCPNet (Yan et al. 2022), Band Aware (Zhou et al. 2023) and MD-CUN (Yang et al. 2022).

Implementation Details

We implement our network on the PC with a single NVIDIA TITAN RTX 3090 GPU, and we build our network in Pytorch framework. The learning rate is set to 1×10^{-4} and the batch size is set to 4.

The frequency of the signal at each input layer is initialized when entering BMHG so that the outputs y_0, y_1, y_2, y_4 are constrained to one eighth, quarter, half, and full bandwidth. Specifically, we set $B_0 = B_1 = B/8$, and $B_2 = B_3 = B_4 = B/4$ such that $\sum_i B_i = B$. We set the number of epochs for each stage of training to 100, 100, 200, and 200, respectively. The experimental results presented in this paper are all testing outputs based on the model of the last training stage.

Datasets and Evaluation Metrics

Datasets. The paired training samples are unavailable in practice. To construct the training dataset, we utilize the Wald protocol (Wald, Ranchin, and Mangolini 1997) to generate the required paired samples. For example, given an origin high-resolution MS image $H \in R^{H\times W\times C}$ and its corresponding PAN image $\bar{P} \in R^{rH\times rW\times c}$ image, both of them are downsampled with ratio r to obtain image pairs $M \in R^{\frac{H}{r}\times\frac{W}{r}\times C}$ and $P \in R^{H\times W\times c}$ in the training set. We use three satellite image datasets in our experiments, including WorldView-II, WorldView-III and GaoFen2, each containing several hundred PAN-LRMS image pairs. The PAN images are cropped into patches with the size of 128×128, and the corresponding LRMS patches are with the size of 32×32.

Evaluation Metrics. We use the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM), SAM, and the relative dimensionless global error in synthesis (ERGAS) as quantitative metrics to evaluate the image quality on three datasets. Additionally, to compare the models' generalization ability, we test them on 200 full-resolution real datasets without down-sampling. The resolution of original PAN images is 128×128, and the corresponding MS images resolution is 32×32. Since this dataset does not contain ground truth, we use three no-reference image quality evaluation metrics to assess model performance, including the spectral distortion index D_λ, the spatial distortion index D_S, and the quality without reference QNR.

Comparison with SOTA Methods

Table 1 presents the performance of our proposed method and the baseline methods on three datasets, with the best results bolded. Our method outperforms existing pan-sharpening methods in all metrics. Specifically, compared to the second-best results, our method achieves a PSNR improvement of 0.13 dB, 0.32 dB, and 1.66 dB on the WorldView-II, WorldView-III and GaoFen2 datasets respectively. Moreover, our method has shown significant improvements over other metrics as well. Additionally, we show the comparison of the visual results in Figure 5 and Figure 6. The top two rows compare the fusion results with SOTA methods, and the bottom row are the MSE residues between the pan-sharpened results and the ground truth. Traditional method (such as GFPCA) tends to lose a significant amount of information during the process of feature dimension reduction, resulting in severe spatial and spectral distortion in the fusion results. By zooming in on the local regions, it is apparent that varying degrees of artifacts present in the results of current deep learning-based methods. This challenge arises due to the difficulty of achieving exact pixel-level alignment within the discrete feature space.

To evaluate the generalization ability of our network, we apply a pre-trained model trained on unseen full-resolution real dataset. Table 2 and Figure 7 show the results. Our method demonstrates superior visual effects in terms of colors and artifacts. For example, the red regions in the bottom-left show clearer contours.

Ablation Experiments

Band-limited Multi-scale High-frequency Generator (BMHG), Spatial-spectral Implicit Image Function (SIIF), and Progressive High-frequency Injection Module (PHIM) are three key modules of our network, we conduct a series of ablation experiments on the WV2 dataset to demonstrate their effectiveness and necessity. We set 5 different configurations for the corresponding network variants of our proposed method and the results of ablation experiments are shown in Table 3.

Band-limited Multi-scale High-frequency Generator. BMHG is used to generate high-quality multi-scale high-frequency signals. For experiment (I) and (II) in Table 3, we

Method	WorldView-II				WorldView-III				GaoFen2			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SFIM	34.1297	0.8975	0.0439	2.3449	21.8212	0.5457	0.1208	8.9730	36.9060	0.8882	0.0318	1.7398
Brovey	35.8646	0.9216	0.0403	1.8238	22.5060	0.5466	0.1159	8.2331	37.7974	0.9026	0.0218	1.3720
GS	35.6376	0.9176	0.0423	1.8774	22.5608	0.5470	0.1217	8.2433	37.2260	0.9034	0.0309	1.6736
IHS	35.2962	0.9027	0.0461	2.0278	22.5579	0.5354	0.1266	8.3616	38.1754	0.9100	0.0243	1.5336
GFPCA	34.5581	0.9038	0.0488	2.1411	22.3344	0.4826	0.1294	8.3964	37.9443	0.9204	0.0314	1.5604
BAM	41.3527	0.9671	0.0239	0.9932	30.3845	0.9188	0.0773	3.1679	45.7419	0.9836	0.0134	0.6267
SFIIN	41.7244	0.9725	0.0220	0.9506	30.5971	0.9236	0.0741	3.0798	47.4712	0.9901	0.0102	0.5462
GCPNet	41.8228	0.9694	0.0227	0.9291	30.5949	0.9227	0.0755	3.0751	47.4165	0.9892	0.0102	0.5472
Band Aware	41.8929	0.9704	0.0223	0.9266	30.6050	0.9220	0.0753	3.0825	47.3186	0.9894	0.0102	0.5508
MDCUN	41.9269	0.9722	0.0215	0.9050	30.5668	0.9227	0.0744	3.0987	47.2023	0.9879	0.0105	0.5533
Ours	**42.0635**	**0.9731**	**0.0210**	**0.8879**	**30.9309**	**0.9264**	**0.0705**	**2.9634**	**49.1388**	**0.9905**	**0.0090**	**0.4469**

Table 1: Quantitative comparison of reference metrics. The best values are bolded. The up or down arrow indicates higher or lower metric corresponding to better images.

Metrics	SFIM	GS	Brovey	IHS	GFPCA	BAM	SFIIN	GCPNet	Band Aware	MDCUN	Ours
$D_\lambda\downarrow$	0.0822	0.0696	0.1378	0.0770	0.0914	0.0755	0.0681	0.0723	0.0701	0.0672	**0.0648**
$D_S\downarrow$	0.1087	0.2456	0.2605	0.2985	0.1635	0.1159	0.1119	0.1144	0.1391	0.1147	**0.1062**
QNR↑	0.8214	0.7025	0.6390	0.6485	0.7615	0.8211	0.8466	0.8265	0.8352	0.8255	**0.8476**

Table 2: Evaluation on the real-world full-resolution scenes from GaoFen2 dataset. The best values are bolded. The up or down arrow indicates higher or lower metric corresponding to better.

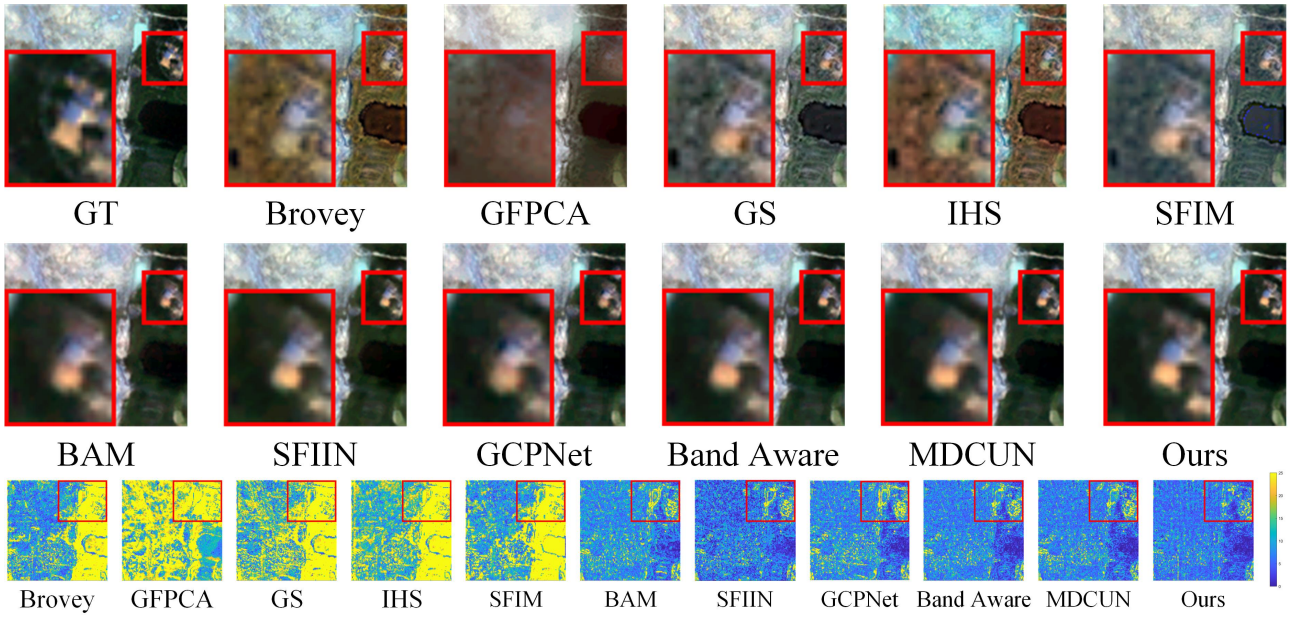


Figure 5: The visual comparisons between other pan-sharpening methods and our method on WorldView-II satellite.

remove the BMHG module and replace it with a dilated network (Fu et al. 2020) to verify its necessity. Table 3 shows that removing BMHG will degrade all metrics. Therefore, BMHG plays a significant role in our network.

Spatial-spectral Implicit Image Function. SIIF is utilized to fuse spatial and spectral features in the continuous domain. For experiment (I) and (III) in Table 3, we validate its effectiveness by replacing SIIF with a CNN (Yang et al. 2017). The results in Table 3 indicate that substituting SIIF with CNN significantly reduces network performance. Therefore, SIIF is crucial in our network.

Progressive High-frequency Injection Module. PHIM enhances the network's ability to fit high-frequency functions by employing a progressive information injection strategy. For experiment (IV) in Table 3, we retain both BMHG and SIIF while removing PHIM. During the training pro-

Config	BMHG	SIIF	PHIM	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	×	×	×	40.9251	0.9661	0.0230	0.9918
(II)	✓	×	×	41.3760	0.9677	0.0221	0.9487
(III)	×	✓	×	41.7689	0.9699	0.0213	0.9111
(IV)	✓	✓	×	41.9474	0.9708	0.0211	0.8946
Ours	✓	✓	✓	**42.0635**	**0.9731**	**0.0210**	**0.8879**

Table 3: Ablation studies about three modules on WV2 dataset. The best values are bolded.

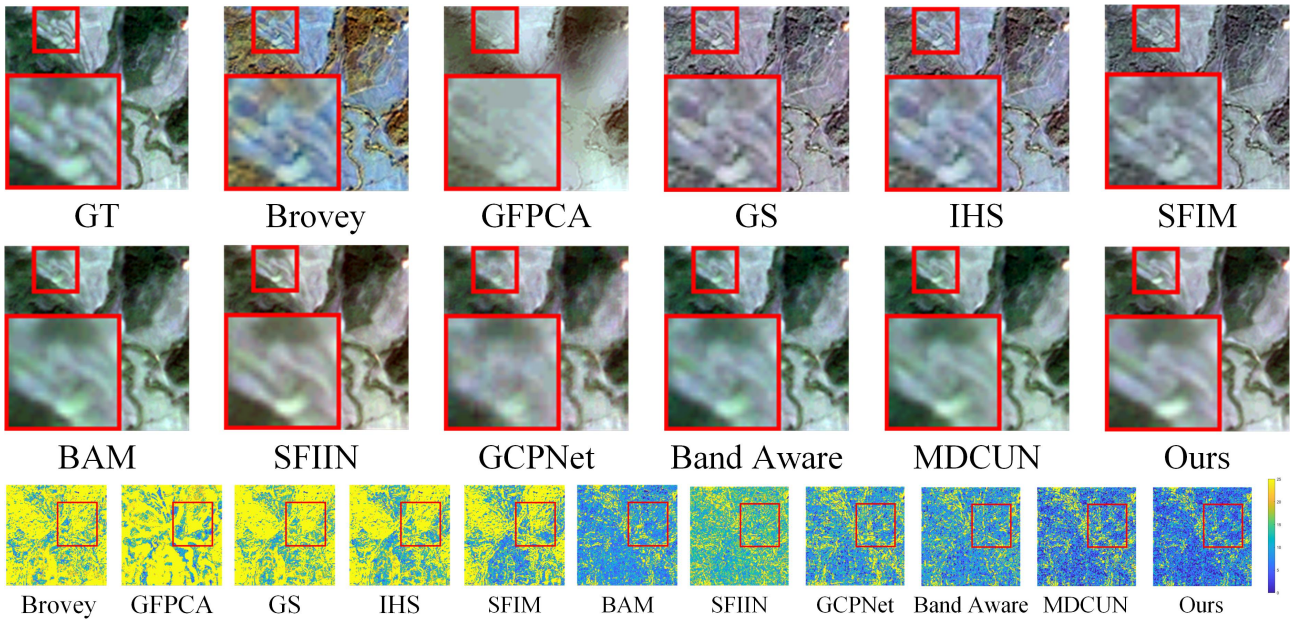


Figure 6: The visual comparisons between other pan-sharpening methods and our method on Gaofen2 satellite.

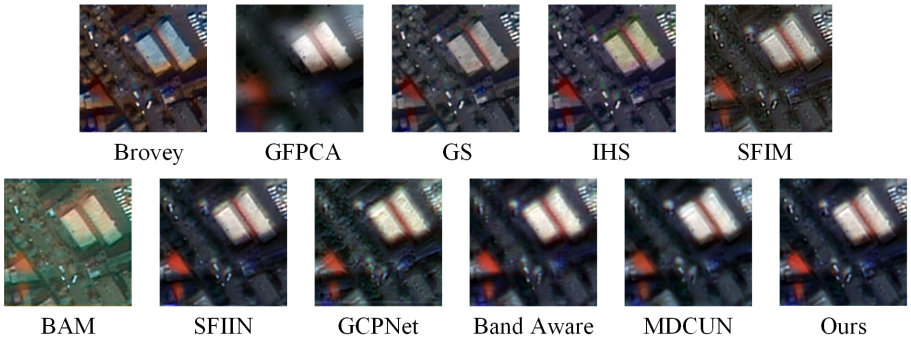


Figure 7: The visual comparisons between other pan-sharpening methods and our method on full-resolution real dataset.

cess, only the full-bandwidth high-frequency signals from the PAN image are fused. The results in Table 3 indicate that including PHIM results in a 0.11dB improvement in PSNR, demonstrating the effectiveness of this module.

Conclusion

In this paper, we propose a novel pan-sharpening network named PIF-Net that aims to progressively restore high-frequency signals of varying scales in the MS image in the implicit space. We design a multi-stage band-limited signal generation network to generate multi-scale high-frequency information and an implicit image representation function is designed to fuse the spatial and spectral features in the continuous domain. Extensive experiments on three satellite datasets demonstrate that our proposed method outperforms existing methods in both quantitative and qualitative metrics.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China under Grant 82172033, U19B2031, 61971369, 52105126, 82272071, 62271430.

References

Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; and Selva, M. 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5): 591–596.

Carper, W.; Lillesand, T.; Kiefer, R.; et al. 1990. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing*, 56(4): 459–467.

Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8628–8638.

Dai, X.; Chen, D.; Liu, M.; Chen, Y.; and Yuan, L. 2020. DA-NAS: Data Adapted Pruning for Efficient Neural Architecture Search.

Fathony, R.; Sahu, A. K.; Willmott, D.; and Kolter, J. Z. 2020. Multiplicative filter networks. In *International Conference on Learning Representations*.

Fu, X.; Wang, W.; Huang, Y.; Ding, X.; and Paisley, J. 2020. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5): 2090–2104.

Gillespie, A. R.; Kahle, A. B.; and Walker, R. E. 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3): 343–365.

Guo, S.; Huang, W.; Haozhi, Z.; Zhuang, C.; Dengke, D.; Scott, M.; and Dinglong, H. 2018. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images.

Guo, Y.; Chen, Y.; Zheng, Y.; Zhao, P.; Chen, J.; Huang, J.; and Tan, M. 2020. Breaking the Curse of Space Explosion: Towards Efficient NAS with Curriculum Search.

Jiang, C.; Sud, A.; Makadia, A.; Huang, J.; Nießner, M.; Funkhouser, T.; et al. 2020. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6001–6010.

Jin, Z.-R.; Deng, L.-J.; Zhang, T.-J.; and Jin, X.-X. 2021. BAM. In *Proceedings of the 29th ACM International Conference on Multimedia*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation.

Khan, M. M.; Chanussot, J.; Condat, L.; and Montanvert, A. 2008. Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. *IEEE Geoscience and Remote Sensing Letters*, 5(1): 98–102.

Laben, C. A.; and Brower, B. V. 2000. Process for enhancing the spatial resolution of multispectral imagery using pansharpening. US Patent 6,011,875.

Liao, W.; Huang, X.; Van Coillie, F.; Thoonen, G.; Pižurica, A.; Scheunders, P.; and Philips, W. 2015. Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter. In *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. Ieee.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Lindell, D. B.; Van Veen, D.; Park, J. J.; and Wetzstein, G. 2022. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16252–16262.

Liu, J. G. 2002. Smoothing Filter-based Intensity Modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3504–3515.

Otazu, X.; González-Audícana, M.; Fors, O.; and Núñez, J. 2005. Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43(10): 2376–2385.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.; Bengio, Y.; and Courville, A. 2019. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 5301–5310. PMLR.

Ranchin, T.; and Wald, L. 2000. Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation. *Photogrammetric engineering and remote sensing*, 66(1): 49–61.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2304–2314.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33: 7462–7473.

Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2022. Curriculum Learning: A Survey. *International Journal of Computer Vision*, 1526–1565.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images.

Wang, Y.; Lin, Y.; Meng, G.; Fu, Z.; Dong, Y.; Fan, L.; Yu, H.; Ding, X.; and Huang, Y. 2023. Learning High-frequency Feature Enhancement and Alignment for Pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, 358–367.

Xu, Z.-Q. J.; Zhang, Y.; and Luo, T. 2022. Overview frequency principle/spectral bias in deep learning. arXiv:2201.07395.

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

Xu, Z.-Q. J.; Zhang, Y.; and Xiao, Y. 2019. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, 264–274. Springer.

Yan, K.; Zhou, M.; Liu, L.; Xie, C.; and Hong, D. 2022. When pansharpening meets graph convolution network and knowledge distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Yang, G.; Zhou, M.; Yan, K.; Liu, A.; Fu, X.; and Wang, F. 2022. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1788–1797.

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

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

Zhou, M.; Yan, K.; Fu, X.; Liu, A.; and Xie, C. 2023. PAN-Guided Band-Aware Multi-Spectral Feature Enhancement for Pan-Sharpening. *IEEE Transactions on Computational Imaging*, 9: 238–249.

Zhou, T.; Wang, S.; and Bilmes, J. 2021. Robust Curriculum Learning: from clean label detection to noisy label self-correction.