Bidomain Modeling Paradigm for Pansharpening

Junming Hou[*]
School of Information Science and Engineering, Southeast University
Nanjing, China
junming_hou@seu.edu.cn

Qi Cao[*]
Yingcai Honors College, University of Electronic Science and Technology of China
Chengdu, China
2020080601007@std.uestc.edu.cn

Ran Ran
School of Mathematical Sciences, University of Electronic Science and Technology of China
Chengdu, China
ranran@std.uestc.edu.cn

Che Liu
School of Information Science and Engineering, Southeast University
Nanjing, China
cheliu@seu.edu.cn

Junling Li[†]
School of Information Science and Engineering, Southeast University
Nanjing, China
junlingli@seu.edu.cn

Liang-jian Deng[†]
School of Mathematical Sciences, University of Electronic Science and Technology of China
Chengdu, China
liangjian.deng@uestc.edu.cn

ABSTRACT

Pansharpening is a challenging low-level vision task whose aim is to learn the complementary representation between spectral information and spatial detail. Despite the remarkable progress, existing deep neural network (DNN) based pansharpening algorithms are still confronted with common limitations. *1)* These methods rarely consider the local specificity of different spectral bands; *2)* They often extract the global detail in the spatial domain, which ignore the task-related degradation, *e.g.,* the down-sampling process of MS image, and also suffer from limited receptive field. In this work, we propose a novel bidomain modeling paradigm for pansharpening problem (dubbed BiMPan), which takes into both local spectral specificity and global spatial detail. More specifically, we first customize the specialized source-discriminative adaptive convolution (SDAConv) for every spectral band instead of sharing the identical kernels across all bands like prior works. Then, we devise a novel Fourier global modeling module (FGMM), which is capable of embracing global information while benefiting the disentanglement of image degradation. By integrating the band-aware local feature and Fourier global detail from these two functional designs, we can fuse a texture-rich while visually pleasing high-resolution MS image. Extensive experiments demonstrate that the proposed framework achieves favorable performance against current state-of-the-art pansharpening methods. The code is available at https://github.com/coder-qicao/BiMPan.

CCS CONCEPTS

• **Computing methodologies → Computer vision problems**.

[*]Authors contributed equally to this research.
[†]Corresponding authors.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612188

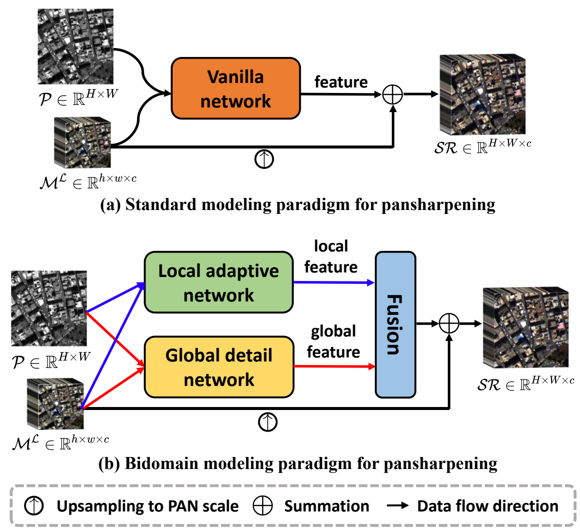

Figure 1: (a) Illustration of commonly used network design, which only works in the spatial domain. (b) Proposed bidomain network scheme, in which the *blue arrow* and *red arrow* denote the information flow in the spatial domain and the Fourier domain, respectively.

KEYWORDS

Pansharpening; Deep neural network; Bidomain modeling; Source-discriminative adaptive convolution; Fourier global modeling

ACM Reference Format:
Junming Hou, Qi Cao, Ran Ran, Che Liu, Junling Li, and Liang-jian Deng. 2023. Bidomain Modeling Paradigm for Pansharpening. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612188

1 INTRODUCTION

The rapid development of satellite sensors have promoted the widespread applications of multispectral (MS) images, such as military system, change detection, and mapping services [1–3]. Notably, both high-resolution spatial details and rich spectral information with respect to MS images are desired in practical applications. Nevertheless, existing remote sensors, such as World-View3 (WV3) and QuickBird (QB), cannot directly capture high spatial resolution MS images due to their physical limitations. Instead, they often observe paired low-resolution MS images and high-resolution panchromatic (PAN) images. Therefore, pansharpening technique is developed to produce high-resolution MS images by super-resolving the low-resolution MS images in the spatial domain, conditioning on the paired PAN images. In other words, pansharpening attempts to borrow the spatial information from PAN images to enhance the spatial resolution of the MS images [4, 5].

To date, numerous pansharpening methods have been proposed by the research community. They can be roughly divided into model-driven methods (also known as traditional methods), including component substitution (CS)-based methods, multi-resolution analysis (MRA)-based methods, variational optimization (VO)-based methods, and deep neural networks (DNNs)-based methods [1–3]. With the success of deep learning in various levels of visual tasks, such as object detection, image segmentation, and single image supers-resolution, explosive DNN methods mainly based on convolutional neural networks (CNNs) have been proposed for pansharpening. The pioneering DNN-based pansharpening method only consists of a three-layer convolution operation [6], which is inspired by the representative single image super-resolution network SRCNN [7]. Afterward, more complicated network architectures have been designed to improve the non-linear representative capacity of pansharpening [8–11]. Although existing pansharpening methods have achieved remarkable progress, they still suffer from some limitations. *First*, they rarely focus on the local specificity with respect to each spectral band, while the local difference among the bands is obvious according to our observation and should not be ignored. *Second*, most DNN-based pansharpening methods commonly conduct detail extraction in the spatial domain as shown in Fig. 1(a), which inevitably suffers from the limited receptive field due to the attribute of the convolution operation. *Besides*, the down-sampling process of MS images often inevitably leads to the loss of high-frequency information, which is tightly coupled to the frequency domain [4, 12, 13]. Given the above facts, we intend to develop a new modeling framework that can take into account both the local specificity of each spectral band and the global contextual detail, as illustrated in Fig. 1(b).

Our Motivation. We first take a high-resolution example from the 4-band QB dataset to demonstrate the difference and connection of different spectral bands. Fig. 2(a) displays the pixel value of every spectral band of MS image and PAN image, and we can clearly see that the pixel distribution of each spectral band varies greatly. This observation reveals a significant fact associated with the pansharpening problem that the local spatial content of different bands is widely diverse. Although prior works attempt to design the content-adaptive convolution kernels to discriminatively deal with the different regions of the input image, they often share identical adaptive kernels across all bands [14–16], which ignores the local specificity of each band. In addition, the gradient statistics of MS image filtered by the Sobel operator are similar to that of PAN image, as shown in Fig. 2(b). This implies that the correlation

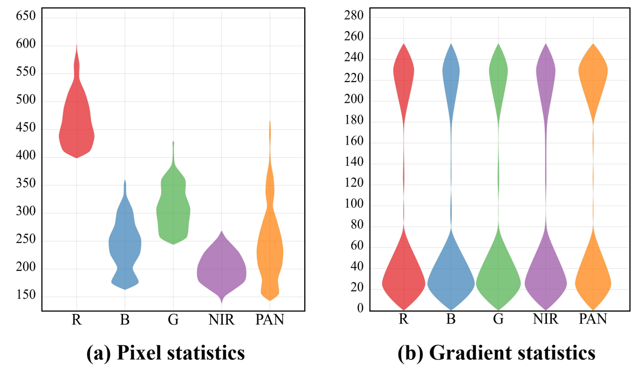

Figure 2: (a) The pixel statistics for every MS band and PAN. (b) The high-frequency gradient statistics for every MS band and PAN. For simplicity, here, we take a 4-band MS image together with paired PAN image as an instance for the purpose of visualization.

between all bands and PAN image is similar in global detail. In other words, an ideally fused high-resolution MS (HRMS) image should be consistent with PAN image in terms of the global detail as much as possible. Currently, pansharpening research community commonly employs multi-scale networks or transformer-based methods to extract the global structure in the spatial domain [17–19]. Nevertheless, they suffer from limited receptive fields and ignore the image priors related to the degradation process.

Based on the above analyses, we consider customizing the specialized adaptive kernels (*i.e.,* source-discriminative adaptive convolution, dubbed as SDAConv) for every band instead of sharing the identical kernels across all bands. The proposed SDAConv is capable of focusing on the local specificity of every band, which is conducive to generating more realistic and content-rich HRMS images. In addition, we intend to extract the global detail in the Fourier domain driven by its nature of global modeling capacity and rich image priors. To be specific, we propose a novel Fourier global modeling module (FGMM) by borrowing the ideas from the existing global modeling paradigm, which neatly incorporates these innate advantages of the Fourier domain into the global modeling rule. By integrating the extracted local features from every spectral band and the global detail, we can predict a desired HRMS image. In conclusion, the contributions of this work can be condensed into the following aspects:

- We propose a novel bidomain modeling paradigm for pansharpening, which achieves the local-global representation learning on HRMS images through two functional designs, *i.e.,* the Band-aware local specificity modeling branch and Fourier global detail reconstruction branch.
- Unlike the prior works, we customize the specialized adaptive convolution kernels for every spectral band given the local differences among various spectral bands, instead of sharing the identical kernels across all band patches. Besides, we propose a novel Fourier global modeling module by borrowing the ideas from the existing global modeling

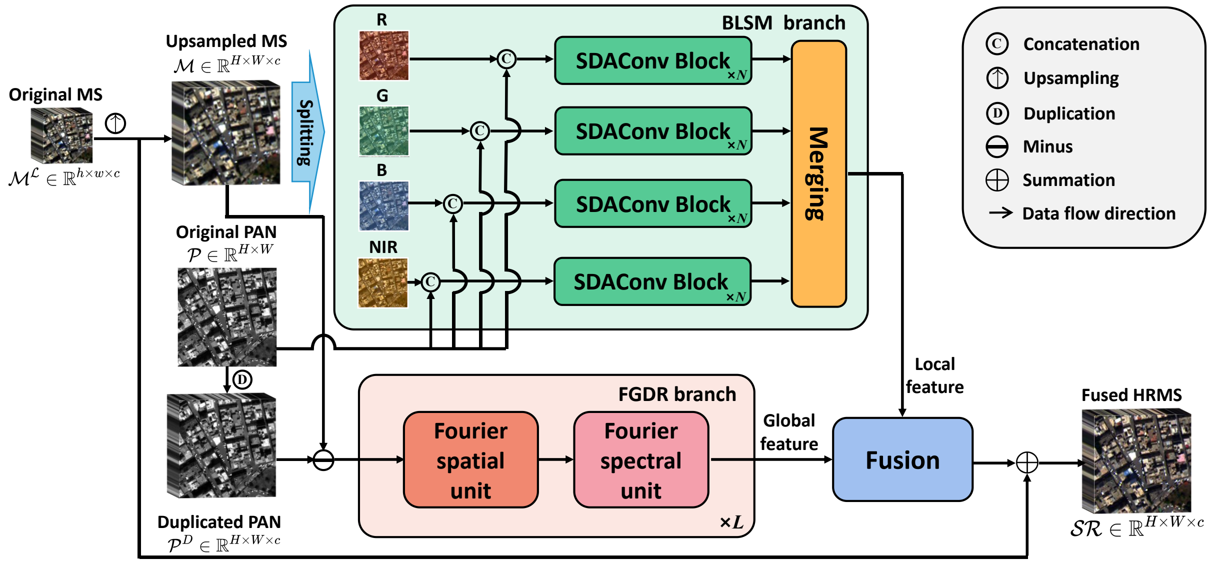

Figure 3: The pipeline of the proposed bidomain modeling framework, which consists of two parts: band-aware local specificity modeling (BLSM) branch and Fourier global detail reconstruction (FGDR) branch. Note that we use a 4-band sample to outline the proposed framework for better illustration.

paradigm, which neatly incorporates the innate advantages of the Fourier domain into the global modeling rule.

- Our proposed modeling framework yields the best qualitative and quantitative results against existing state-of-the-art approaches on different satellite datasets and also generalizes well to real-world full-resolution scenes.

2 RELATED WORKS

Adaptive Convolution Techniques. Standard convolution suffers from its inherent spatial-invariance property, which leads to limited performance in some pixel-level vision tasks, *e.g.*, single image super-resolution, and pansharpening. In recent years, adaptive convolution techniques have aroused much attention in the computer vision research community due to their flexibility, in which the sampling locations and/or kernel values are adjusted according to the input content [20–22]. For pansharpening task, some representative adaptive convolution techniques also have been proposed and shown favorable performance in comparison to standard convolution. In [14], researchers first design a novel adaptive convolution that includes both local content and global harmonic basis, dubbed LAGConv, which can effectively exploit local specificity and integrate global information of the involved image patch. Inspired by the LAGConv, Lu *et. al.* [16] proposes a lightweight pansharpening network consisting of several adaptive feature learning blocks. In [23], the distinctive attributes of input source images are considered to design the so-called source-adaptive discriminative kernels, which consist of two components, *i.e.,* spatial kernels derived from texture-rich PAN images and spectral kernels derived from MS images. To tackle the computational cost of standard convolution operation, Chen *et. al.* [24] devises an interpretable span strategy to generate the convolution kernels, which only learns two navigated kernels, and then extends them to all channels.

Fourier Based DNN Networks. In low-level vision tasks, most existing DNN-based methods are designed to learn the non-linear mapping between the inputs and outputs in the spatial domain, which inevitably suffer from limited receptive fields. Recently, the Fourier domain has gained much attention due to its unique characteristics, e.g., image priors and global modeling attribute. In [25], researchers attempt to address the low-light image enhancement problem in the Fourier domain. Mao *et. al.* [26] embeds a novel Res FFT-ReLU Block into the cascaded network for image deblurring, which learns the spatial-frequency bidomain representations to extract both kernel-level and pixel-level features. Likewise, [27] and [28] adopt a similar methodology to deal with the related issues. In [29], authors first explore the commonly used spatial down-/up-sampling operation from the perspective of the Fourier domain and design a plug-and-play FourierUP operator, which is capable of modeling the global dependency, thus breaking the common limitation of spatial operators. For pansharpening, some representative works have been proposed by combining spatial-frequency dual-domain information to reconstruct HRMS images [4, 13, 30].

3 METHODOLOGY

3.1 Overall Architecture

Our main goal is to explore an effective modeling paradigm for pansharpening based on the aforementioned facts, which can produce a texture-rich while visually pleasing HRMS image. To this end, we attempt to take into account both the local specificity of each spectral band and the global contextual detail. Specifically, we devise two core designs, *i.e.,* band-aware local specificity modeling branch and Fourier global detail reconstruction branch. We first introduce

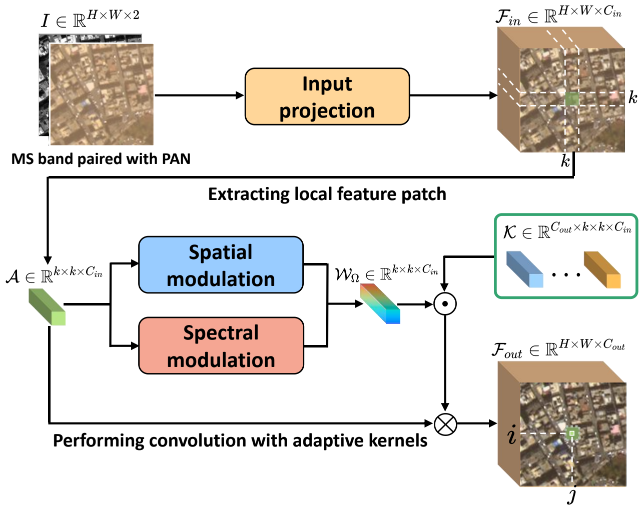

Figure 4: Structure of the proposed source-discriminative adaptive convolution (SDAConv).

the overall pipeline of our framework, and then we elaborate on the detailed structures of two functional branches.

Overall Pipeline. Fig. 3 outlines the overall architecture of our method. For the upper branch, c parallel subbranches are adopted to focus on the local specificity of every band, where c denotes the number of spectral bands. Given an up-sampled MS image $\mathcal{M} \in \mathbb{R}^{H\times W\times c}$, we first split it into c single bands. Then, we send each band paired with a PAN image $\mathcal{P}$ to the corresponding subbranch to extract its local features. Specifically, we first apply a convolution layer as well as the ReLU function to extract the shallow features of the input band and PAN image. After that, we concatenate the obtained band features F_i^{spe} and PAN features F^{spa} as the input data of the corresponding subbranch to extract the local information. Finally, the outputs of all subbranches are integrated to obtain the local features.

In addition, we devise a novel Fourier global modeling module (FGMM) to construct the below branch. For a pair of input images, *i.e.,* the up-sampled MS image $\mathcal{M}$ and PAN image $\mathcal{P}$, we first duplicate $\mathcal{P}$ along the spectral dimension. Then, the high-frequency contents of $\mathcal{M}$ are obtained by subtracting every band from $\mathcal{P}$, which is used as the input data to reconstruct the global detail. Finally, we integrate the obtained local feature and global detail, and add the up-sampled MS images to produce the super-resolution MS images.

3.2 Band-Aware Local Specificity Modeling Branch

Source-Discriminative Adaptive Convolution. Unlike existing adaptive convolution techniques that usually generate the kernels using a certain image patch including all bands. To explore the local uniqueness of each band, we devise a new adaptive convolution operation, denoted as source-discriminative adaptive convolution (SDAConv), whose kernels are generated depending on the local patch of every MS band coupled with PAN image, as shown in Fig. 4. Without loss of generality, we take a single band as an instance to introduce the kernel generation process of the proposed SDAConv. In fact, the kernel generation process aims to obtain a weight matrix $\mathcal{W} \in \mathbb{R}^{k\times k\times C_{in}}$, where k and C_{in} represent the size and the number of channels corresponding to the input local patch, respectively. To be specific, given an input patch $k \times k \times C_{in}$, it is first projected to the low-level features through a layer convolution layer. Next, we apply a spatial modulation and a spectral modulation to deal with the extracted shallow features. Then, we further combine the outputs from these two modules to generate the expected weight matrix $\mathcal{W}_\Omega$, which can be illustrated as follows:

$$\begin{aligned} F_\Omega^{spa}, F_\Omega^{spe} &= SM(\sigma(Conv(A))), CM(\sigma(Conv(A))), \\ \mathcal{W}_\Omega &= Re(F_\Omega^{spa} \odot F_\Omega^{spe}), \end{aligned} \tag{1}$$

where A represents the input local patch, $Conv(\cdot)$ and $\sigma(\cdot)$ are the convolution layer and ReLU non-linear function, $SM(\cdot)$ and $CM(\cdot)$ denote the spatial modulation operation and channel modulation operation, respectively. F_Ω^{spa} and F_Ω^{spe} correspond to the outputs of $SM(\cdot)$ and $CM(\cdot)$. The symbol $\odot$ is the inner product operation and $Re(\cdot)$ represents the reshape operation, while the symbol W_Ω denotes the obtained weight matrix.

After that, we perform this weight matrix $\mathcal{W}_\Omega$ on a group of candidate kernels $\mathcal{K} \in \mathbb{R}^{C_{out}\times k\times k\times C_{in}}$ to generate the adaptive kernels. Finally, we can apply the obtained adaptive convolution kernels to the feature maps of the input patch. Briefly, they can be formulated as follows: $\widetilde{\mathcal{K}}_\Omega \in \mathbb{R}^{C_{out}\times k\times k\times C_{in}}$

$$D_\Omega = \sigma(Conv(A)) \otimes (\mathcal{W}_\Omega \odot \mathcal{K}), \tag{2}$$

where $\otimes$ and D_Ω denote the convolution operation and the corresponding output, respectively.

Band-aware Local Specificity Modeling Branch. Our band-aware local specificity modeling (BLSM) branch consists of c subbranches, where c is the number of the spectral band corresponding to the MS image. All subbranches adopt the same structure that contains several cascaded SDAConv blocks, but does not share the parameters due to the difference of source inputs, *i.e.,* different bands of MS images. In addition, the outcomes of all subbranches are integrated through a merging module. The entire procedure can be mathematically represented as follows:

$$\begin{aligned} O_\lambda^i &= \Phi(Cat(Conv(\mathcal{M}_i), Conv(\mathcal{P}))), \\ F_{lo} &= \Psi(O_\lambda^1, O_\lambda^2, \cdots, O_\lambda^c), \end{aligned} \tag{3}$$

where $Cat(\cdot)$ is concatenation operation, $\Phi(\cdot)$ represents the mapping function of each subbranch, and O_λ^i is the output of the i-th subbranch. The symbol $\Psi(\cdot)$ denotes integrating the outputs from all subbranches, and F_{lo} is the final outcome of the proposed BLSM branch.

3.3 Fourier Global Detail Reconstruction Branch

Extracting accurate global detail is a popular yet challenging issue in pansharpening tasks. Most existing approaches often address this problem in the spatial domain, which ignore the degradation of MS images and require a high computational cost. Recently, the Fourier domain has gained extensive attention from the research community. On the one hand, the Fourier transform is capable of capturing the image priors with respect to the down-sampling

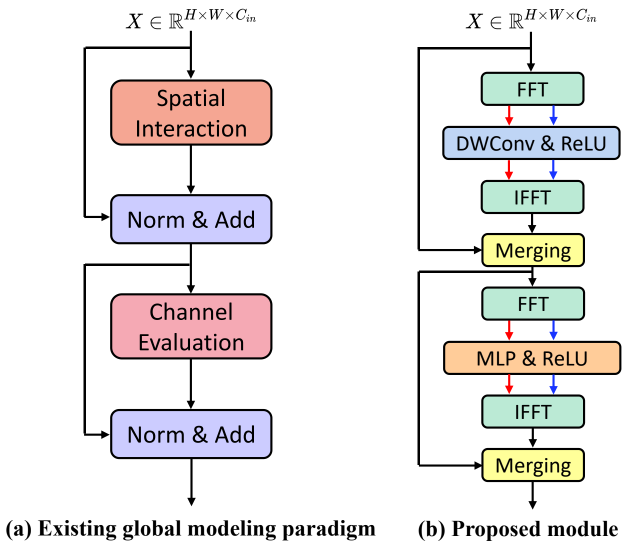

Figure 5: (a) Structure of existing global modeling paradigm. (b) Proposed Fourier global modeling module. Notably, FFT and IFFT represent the Fourier transform and inverse Fourier transform, while the *red arrow* and the *blue arrow* denote the real part and the imaginary part, respectively.

process of MS images. On the other hand, the Fourier domain associates each pixel in such space with all spatial pixels owing to its innate global attributes. Therefore, we attempt to borrow the ideas from existing global modeling paradigms, *e.g.,* [31–33], to explore the analogous design in Fourier domain to reconstruct the global details.

Preliminary. Given an image $x \in \mathbb{R}^{H\times W\times C}$, we can employ the Fourier transformation $\mathcal{F}(\cdot)$ to convert it into a complex component in the Fourier space, which can be mathematically represented as follows:

$$\mathcal{F}(x)(u,v) = \frac{1}{H\times W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \quad (4)$$

where u and v represent the coordinates in Fourier space. In turn, we can use the inverse Fourier transform $\mathcal{F}^{-1}(\cdot)$ to achieve the transformation from the Fourier domain to the spatial domain. Besides, the amplitude and phase of x in Fourier space can be calculated using the following formula:

$$\begin{aligned} A(x)(u,v) &= \sqrt{R^2(x)(u,v) + I^2(x)(u,v)}, \\ P(x)(u,v) &= arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \end{aligned} \quad (5)$$

where *R(x)* and *I(x)* are the real part and the imaginary part, respectively. Notably, both $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ can be independently performed on each channel of the input data.

Structure of the Fourier Global Modeling Module. Fig. 5 gives the detailed architecture of the proposed Fourier global modeling module (FGMM), which consists of two components, *i.e.,* Fourier spatial unit and Fourier spectral unit.

Fourier Spatial Unit. Given an input feature $X \in \mathbb{R}^{H\times W\times C_{in}}$, it will go through two paths in parallel, *i.e.,* Fourier operation and spatial operation. In terms of the former, the Fourier transform is first conducted as follows:

$$X_R,\ X_I = \mathcal{F}(X), \quad (6)$$

where X_R and X_I represent the real and imaginary parts, respectively. Afterward, we adopt a 3×3 depth-wise convolution layer coupled with the ReLU activation function to integrate the spatial information, which can be expressed as follows:

$$S_R,\ S_I = \sigma(DW(X_R)),\ \sigma(DW(X_I)), \quad (7)$$

where $DW(\cdot)$ represents the depth-wise convolution. Then, we transform the obtained S_R and S_I back to the spatial domain through the inverse Fourier transform, illustrated as follows:

$$Z_F = \mathcal{F}^{-1}(S_R, S_I). \quad (8)$$

Besides, we also adopt a spatial path to complement the spatial structure information, in which the input feature is directly fed into a 3×3 depth-wise convolution layer followed by the ReLU activation function, which can be written as follows:

$$Z_S = \sigma(DW(X)), \quad (9)$$

Next, we fuse the Fourier features Z_F and spatial features Z_S through an efficient Half Instance Normalization (HIN) block[34], which is represented as follows:

$$U = H_S(Z_F, Z_S), \quad (10)$$

where $H_S(\cdot)$ represents integrating the spatial information using an HIN block and U is the output features.

Fourier Spectral Unit. Similar to the Fourier spatial integration, we take an analogous design, *i.e.,* Fourier modulation path and spatial modulation path, to implement the Fourier spectral adjustment, in which the 3×3 depth-wise convolution is replaced by point-wise convolution. Firstly, the Fourier transform is utilized to decompose the output U from the Fourier spatial integration into real and imaginary components, *i.e.,* U_R and U_I. Then, we adopt a point-wise convolution layer followed by ReLU non-linear activation to perform the Fourier channel adjustment. Specifically, this procedure can be formulated as follows:

$$\begin{aligned} &C_R,\ C_I = \sigma(MLP(\mathcal{F}(U_R))),\ \sigma(MLP(\mathcal{F}(U_I))), \\ &C_F = \mathcal{F}^{-1}(C_R, C_I), \end{aligned} \quad (11)$$

where $MLP(\cdot)$ represents the point-wise convolution and C_F is the output of the Fourier channel adjustment.
Afterward, we integrate the outcomes from the frequency path and spatial path via the HIN block to output the final result of our Fourier global modeling module (FGMM), which can be mathematically written as follows:

$$\begin{aligned} C_S &= \sigma(MLP(U)), \\ Out &= H_C(C_F, C_S), \end{aligned} \quad (12)$$

where C_S is the output from the spatial channel modulation. $H_C(\cdot)$ denotes adjusting the channel information using an HIN block, while *Out* is the outcome of the proposed FGMM.

Table 1: Average quantitative metrics on 20 reduced-resolution and 20 full-resolution samples for the WV3 dataset. Some traditional methods (the first four rows) and CNN methods are compared. (Bold: best; Underline: second best)

method	Reduced				Full		
	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(± std)	D_λ(± std)	D_s(± std)	QNR(± std)
BDSD-PC	5.4675±1.7185	4.6549±1.4667	0.8117±0.1063	0.9049±0.0419	0.0231±0.0171	0.0730±0.0356	0.9061±0.0474
MTF-GLP-FS	5.3233±1.6548	4.6452±1.4441	0.8177±0.1014	0.8984±0.0466	0.0354±0.0211	0.0630±0.0284	0.9043±0.0454
BT-H	4.8985±1.3028	4.5150±1.3315	0.8182±0.1019	0.9240±0.0243	0.0430±0.0232	0.0810±0.0374	0.8803±0.0540
CDIF	4.8548±1.4788	4.5029±1.5338	0.8322±0.1032	0.9163±0.0298	0.0317±0.0075	0.0305±0.0152	0.9389±0.0213
PNN	3.6798±0.7625	2.6819±0.6475	0.8929±0.0923	0.9761±0.0075	0.0213±0.0080	0.0428±0.0147	0.9369±0.0212
DiCNN	3.5929±0.7623	2.6733±0.6627	0.9004±0.0871	0.9763±0.0072	0.0362±0.0111	0.0462±0.0175	0.9195±0.0258
MSDCNN	3.7773±0.8032	2.7608±0.6884	0.8900±0.0900	0.9741±0.0076	0.0230±0.0091	0.0467±0.0199	0.9316±0.0271
BDPN	4.1646±0.8223	3.0335±0.7269	0.8724±0.0979	0.9677±0.0087	0.0395±0.0251	0.0459±0.0187	0.9168±0.0404
FusionNet	3.3252±0.6978	2.4666±0.6446	0.9044±0.0904	0.9807±0.0069	0.0239±0.0090	0.0364±0.0137	0.9406±0.0197
LagNet	3.1042±0.5585	2.2999±0.6128	0.9098±0.0907	0.9838±0.0068	0.0368±0.0148	0.0418±0.0152	0.9230±0.0247
BiMPan(ours)	**2.9842±0.6009**	**2.2569±0.5520**	**0.9153±0.0865**	**0.9843±0.0049**	**0.0170±0.0128**	**0.0344±0.0144**	**0.9493±0.0255**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

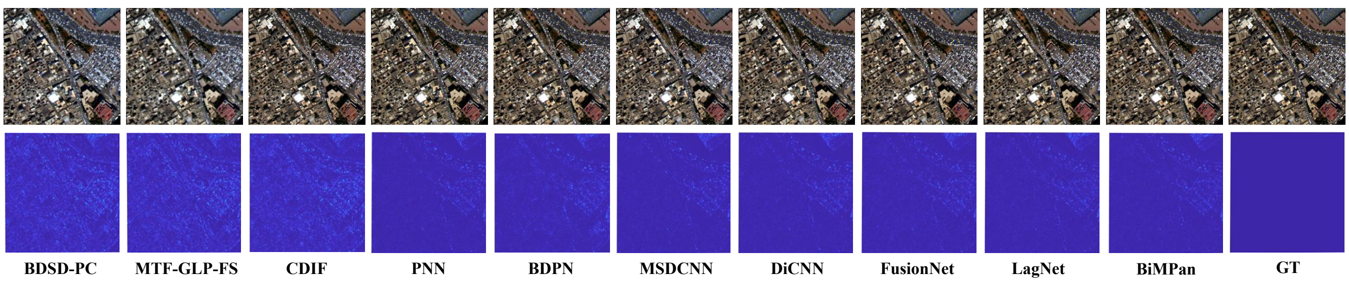

Figure 6: Qualitative comparison on the reduced-resolution sample from WV3 dataset. The first row demonstrates the RGB visualization, while the corresponding absolute error maps are presented in the second row.

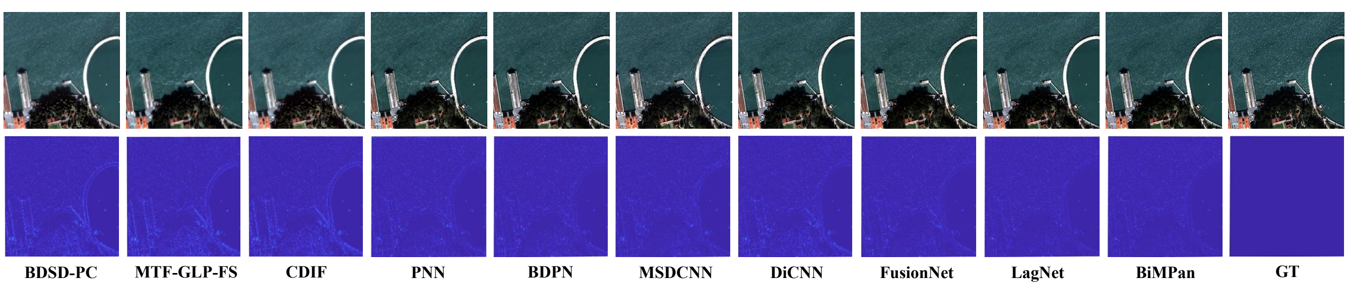

Figure 7: Qualitative comparison on the reduced-resolution sample from QB dataset. The first row demonstrates the RGB visualization, while the corresponding absolute error maps are presented in the second row.

4 EXPERIMENTS

Due to the page limitation, experiment settings including datasets, metrics, benchmarks, and implementation details are provided in the appendix.

4.1 Evaluation on Reduced-Resolution

The reduced-resolution evaluation is conducted to measure the difference between the predicted SR images and the GT images. Quantitative results of all compared methods and our model on 20 WV3 testing examples are presented in Table 1. It is clearly seen that our model achieves the best performance on all indexes, which well demonstrates the superiority of the proposed method. On the one hand, our network can focus on the local specificity of each band through the customized source-discriminative adaptive convolution; on the other hand, it is capable of effectively extracting the global details using the natural advantages of the Fourier domain in global modeling. Fig. 6 presents the visual comparisons among all compared pansharpening methods. As shown in the figures, the

Table 2: Average quantitative metrics on 20 reduced-resolution and 20 full-resolution samples for the QB dataset. Some traditional methods (the first four rows) and CNN methods are compared. (Bold: best; Underline: second best)

method	Reduced				Full		
	SAM(± std)	ERGAS(± std)	Q4(± std)	SCC(± std)	D_λ(± std)	D_s(± std)	QNR(± std)
BDSD-PC	8.2620±2.0497	7.5420±0.8138	0.8323±0.1013	0.9030±0.0181	0.0345±0.0172	0.1636±0.0483	0.8078±0.0497
MTF-GLP-FS	8.1131±1.9553	7.5102±0.7926	0.8296±0.0905	0.8998±0.0196	0.0570±0.0137	0.1500±0.0238	0.8017±0.0295
BT-H	7.1943±1.5523	7.4008±0.8378	0.8326±0.0880	0.9156±0.0152	0.0526±0.0141	0.1648±0.0167	0.7912±0.0177
CDIF	7.2961±1.6703	7.1086±0.7077	0.8460±0.0918	0.9118±0.0139	**0.0175±0.0137**	<u>0.0486±0.0298</u>	<u>0.9351±0.0398</u>
PNN	5.2054±0.9625	4.4722±0.3734	0.9180±0.0938	0.9711±0.0123	0.0569±0.0112	0.0624±0.0239	0.8844±0.0304
DiCNN	5.3795±1.0266	5.1354±0.4876	0.9042±0.0942	0.9621±0.0133	0.0920±0.0143	0.1067±0.0210	0.8114±0.0310
MSDCNN	5.1471±0.9342	4.3828±0.3400	0.9176±0.0987	0.9722±0.0124	0.0320±0.0237	0.0667±0.0282	0.9041±0.0466
BDPN	6.1225±1.2106	5.2756±0.6870	0.8991±0.0938	0.9580±0.0154	0.0734±0.0273	0.0492±0.0126	0.8812±0.0336
FusionNet	4.9226±0.9077	4.1594±0.3212	0.9252±0.0902	0.9755±0.0104	0.0586±0.0189	0.0522±0.0088	0.8922±0.0219
LagNet	**4.5548±0.8155**	<u>3.8436±0.4032</u>	<u>0.9303±0.0935</u>	**0.9805±0.0088**	0.0844±0.0238	0.0676±0.0136	0.8536±0.0178
BiMPan(ours)	<u>4.5860±0.8206</u>	**3.8394±0.3187**	**0.9311±0.0908**	<u>0.9800±0.0079</u>	<u>0.0259±0.0201</u>	**0.0399±0.0121**	**0.9355±0.0298**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

Table 3: Average quantitative metrics on 20 reduced-resolution and 20 full-resolution samples for the WV2 dataset. Some CNN methods are compared. (Bold: best; Underline: second best)

method	Reduced				Full		
	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(± std)	D_λ(± std)	D_s(± std)	QNR(± std)
PNN	7.1158±1.6812	5.6152±0.9431	0.7619±0.0928	0.8782±0.0175	0.1484±0.0957	0.0771±0.0169	0.7869±0.0959
DiCNN	6.9216±0.7898	6.2507±0.5745	0.7205±0.0746	0.8552±0.0289	0.1412±0.0661	0.1023±0.0195	0.7700±0.0505
MSDCNN	<u>6.0064±0.6377</u>	<u>4.7438±0.4939</u>	<u>0.8241±0.0799</u>	0.8972±0.0109	0.0589±0.0421	<u>0.0290±0.0138</u>	<u>0.9143±0.0516</u>
BDPN	7.0934±0.8630	4.8568±0.5698	0.8235±0.0929	0.9033±0.0094	0.1117±0.0859	0.0328±0.0243	0.8606±0.0979
FusionNet	6.4257±0.8602	5.1363±0.5151	0.7961±0.0737	0.8746±0.0134	<u>0.0519±0.0292</u>	0.0559±0.0146	0.8948±0.0187
LagNet	6.9545±0.4739	5.3262±0.3185	0.8054±0.0837	<u>0.9125±0.0101</u>	0.1302±0.0856	0.0547±0.0159	0.8229±0.0884
BiMPan(ours)	**5.7496±0.6008**	**4.5111±0.4837**	**0.8271±0.1043**	**0.9127±0.0089**	**0.0184±0.0064**	**0.0386±0.0137**	**0.9438±0.0181**
Ideal Value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

fused result from our model is very close to the GT image. Furthermore, we also conduct experiments on a 4-band dataset QB, to validate the wide applicability of the proposed method. Table 2 shows the outcomes for all baselines. As expected, our model outperforms other compared pansharpening approaches on the 4-band dataset as well, which further proves the effectiveness of the proposed method. The visual comparison results are provided in Fig. 7.

4.2 Evaluation on Full-Resolution

The goal of pansharpening is to achieve real-world applications. Therefore, we also conduct the full-resolution analysis to corroborate the generalization capability of the reduced-resolution results, in which the GT images are unavailable. Specifically, we use 20 WV3 and 20 QB images, respectively, to perform the full-resolution experiments. Table 1 presents the quantitative comparisons of all compared methods on the WV3 dataset. Again, the proposed method yields the best results. Fig. 8 displays the qualitative comparisons on WV3, in which our model produces visual fidelity images against other methods. Considering the 4-band QB dataset, our method also outperforms other benchmarks in both quantitative and qualitative results at full-resolution, as demonstrated in Table 2 and Fig. 9, respectively.

4.3 Generalization Capability

To further verify the generalization ability of the proposed approach, we directly use 20 WV2 samples from the reduced-resolution to test all DNN-based models that are trained on the WV3 dataset. Table 3 demonstrates the quantitative evaluation results, from which our method performs the favorable generalization capability in comparison to other DNN-based pansharpening techniques.

4.4 Ablation Study

In this section, we conduct some ablation experiments to prove the effectiveness of the designed components, as well as the universality of the proposed framework.

Different Convolution Operations. We compare the performance of different types of convolution operation, including standard convolution, involution operator, and the proposed SDAConv. Table 4 indicates that both adaptive convolutions outperform the standard convolution, while our adaptive technique is superior to the involution operator.

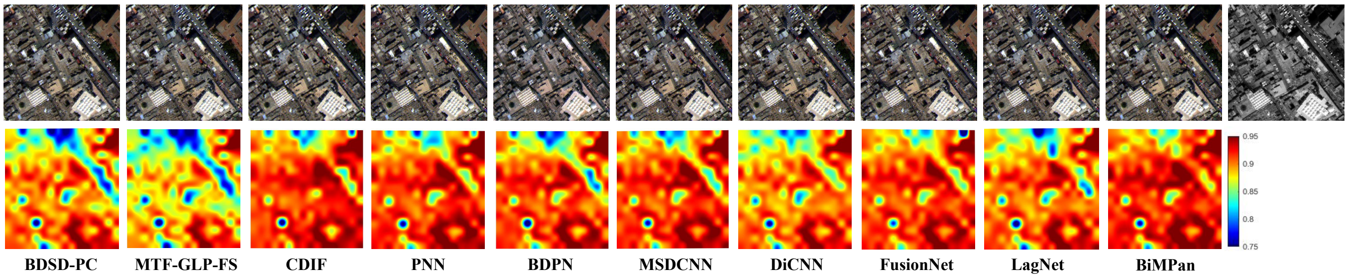

Figure 8: Qualitative comparison on the full-resolution sample from WV3 dataset. The first row presents the RGB visualization, while the second row gives the corresponding QNR maps. The rightmost image is PAN.

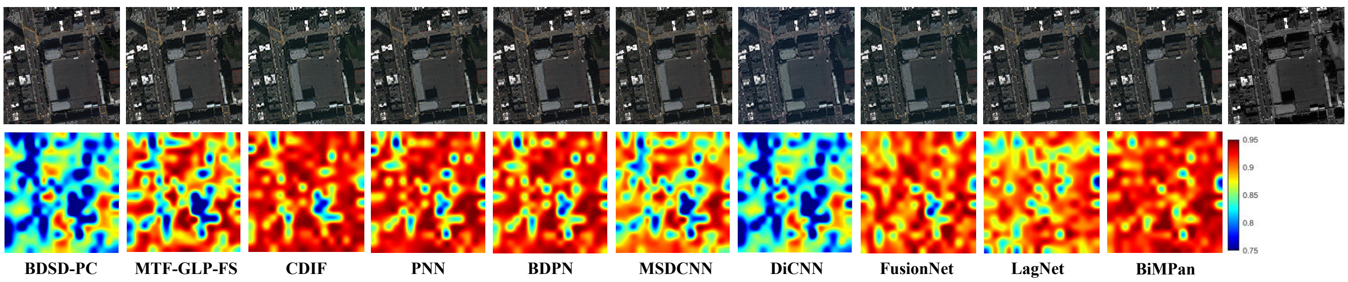

Figure 9: Qualitative comparison on the reduced-resolution sample from QB dataset. The first row demonstrates the RGB visualization, while the second row gives the corresponding QNR maps. The rightmost image is PAN.

Different Strategies of Global Detail Extraction. We conduct

Table 4: Qualitative comparisons of the different convolution operations.

Methods	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(± std)
Standard convolution	3.30±0.68	2.47±0.67	0.91±0.09	**0.98±0.01**
Involution	3.04±0.61	2.28±0.58	**0.92±0.09**	**0.98±0.01**
SDAConv	**2.98±0.60**	**2.26±0.55**	**0.92±0.09**	**0.98±0.01**

the global detail extraction through two strategies, *i.e.,* using the proposed FGMM and FGMM without the Fourier transform (w/o FFT). From Table 5, we can observe that our FGMM achieves better outcomes due to its comprehensive superiority.

Table 5: Qualitative comparisons of strategies of global details extraction.

Methods	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(± std)
w/o FFT	3.16±0.61	2.46±0.63	0.91±0.09	**0.98±0.01**
FGMM	**2.98±0.60**	**2.26±0.55**	**0.92±0.09**	**0.98±0.01**

Different Types of Input Sources for Extracting Global Details. We also investigate the effects of different input sources: 1) MS images and PAN images, denoted as general inputs; 2) the high-frequency components of MS images obtained by subtracting each band from PAN, represented as HFC. From Table 6, we can see that using high-frequency components as input data obtains favorable results.

Table 6: Qualitative comparisons of strategies of global details extraction.

Input sources	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(± std)
General inputs	3.21±0.65	2.43±0.63	0.92±0.09	**0.98±0.01**
HFC	**2.98±0.60**	**2.26±0.55**	**0.92±0.09**	**0.98±0.01**

5 CONCLUSION

We propose a novel bidomain modeling framework for pansharpening, which consists of a band-aware local specificity modeling branch and a Fourier global detail reconstruction branch. Specifically, the former is utilized to focus on the local specificity of each spectral band through the customized source-discriminative adaptive convolution. While the latter is devised to extract the global detail using the innate properties of the Fourier domain, *e.g.,* the disentanglement of image degradation and global modeling capability. By integrating the complementary information from the well-designed two branches, our model outperforms the existing state-of-the-art pansharpening methods on a wide range of benchmark datasets. Specially, it is capable of generalizing well to real-world full-resolution scenes.

ACKNOWLEDGMENTS

The work is supported by National Natural Science Foundation of China (Grant No. 12271083), and Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0501) and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515110949).

REFERENCES

[1] Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus O Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81, 2020.

[2] Xiangchao Meng, Yiming Xiong, Feng Shao, Huanfeng Shen, Weiwei Sun, Gang Yang, Qiangqiang Yuan, Randi Fu, and Hongyan Zhang. A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):18–52, 2020.

[3] Liang-Jian Deng, Gemine Vivone, Mercedes E Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022.

[4] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pansharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 274–291. Springer, 2022.

[5] Cheng Jin, Liang-Jian Deng, Ting-Zhu Huang, and Gemine Vivone. Laplacian pyramid networks: A new approach for multispectral pansharpening. *Information Fusion*, 78:158–170, 2022.

[6] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[8] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.

[9] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1188–1204, 2019.

[10] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2090–2104, 2020.

[11] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2020.

[12] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2360–2369, 2021.

[13] Man Zhou, Jie Huang, Chongyi Li, Hu Yu, Keyu Yan, Naishan Zheng, and Feng Zhao. Adaptively learning low-high frequency information integration for pansharpening. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3375–3384, 2022.

[14] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1113–1121, 2022.

[15] Kaiwen Zheng, Jie Huang, Man Zhou, Danfeng Hong, and Feng Zhao. Deep adaptive pansharpening via uncertainty-aware image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[16] Hangyuan Lu, Yong Yang, Shuying Huang, Xiaolong Chen, Biwei Chi, Aizhu Liu, and Wei Tu. Awfln: An adaptive weighted feature learning network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[17] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei 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*, 11(3):978–989, 2018.

[18] Man Zhou, Jie Huang, Xueyang Fu, Feng Zhao, and Danfeng Hong. Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[19] Wele Gedara Chaminda Bandara and Vishal M Patel. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1767–1777, 2022.

[20] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019.

[21] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021.

[22] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12321–12330, 2021.

[23] Siran Peng, Liang-Jian Deng, Jin-Fan Hu, and Yuwei Zhuo. Source-adaptive discriminative kernels based network for remote sensing pansharpening. In *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022.

[24] Zhi-Xuan Chen, Cheng Jin, Tian-Jing Zhang, Xiao Wu, and Liang-Jian Deng. Spanconv: A new convolution via spanning kernel space for lightweight pansharpening. In *Proc. 31st Int. Joint Conf. Artif. Intell.*, pages 1–7, 2022.

[25] Chongyi Li, Chun-Le Guo, man zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *The Eleventh International Conference on Learning Representations*, 2023.

[26] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023.

[27] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 181–198. Springer, 2022.

[28] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 163–180. Springer, 2022.

[29] man zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[30] Zheng Wang, Yanwei Zhao, and Jiacheng Chen. Multi-scale fast fourier transform based attention network for remote-sensing image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2728–2740, 2023.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[33] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[34] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021.

[35] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992.

[36] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002.

[37] Jie Zhou, Daniel L Civco, and John A Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International journal of remote sensing*, 19(4):743–757, 1998.

[38] Andrea Garzelli and Filippo Nencini. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):662–665, 2009.

[39] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2014.

[40] Simone Lolli, Luciano Alparone, Andrea Garzelli, and Gemine Vivone. Haze correction for contrast-based multispectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2255–2259, 2017.

[41] Gemine Vivone. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing*, 57(9):6421–6433, 2019.

[42] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing*, 27(7):3418–3431, 2018.

[43] Jin-Liang Xiao, Ting-Zhu Huang, Liang-Jian Deng, Zhong-Cheng Wu, and Gemine Vivone. A new context-aware details injection fidelity with adaptive coefficients estimation for variational pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

In this Appendix, we present the loss function, detailed experiment settings, and additional experimental results.

A LOSS FUNCTION

For simplicity, we choose the $\mathcal{L}_1$ loss function to minimize the difference between the predicted super-resolution images SR and the ground truth images GT during the network training process, which can be represented as follows:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|SR_i - GT_i\|_1 , \tag{13}$$

where N denotes the number of training samples, and $\|\cdot\|_1$ is the $\mathcal{L}_1$ norm.

B EXPERIMENT SETTINGS

Datasets. We investigate the effectiveness of the proposed method on a wide range of datasets, including 8-band datasets from WorldView-3 (WV3) and WorldView2 (WV2) sensors, and 4-band datasets from QuickBird (QB) sensors. Notably, we leverage Wald's protocol to simulate the source data due to the unavailability of ground truth (GT) images. All training data (*i.e.,* PanCollection dataset[3]) used in this work is available on the public website (https://liangjiandeng.github.io/PanCollection.html.), which includes fair and detailed data description. Take WV3 as an instance, we use 10000 PAN ($64 \times 64 \times 1$)/LRMS ($64 \times 64 \times 8$)/GT($64 \times 64 \times 8$) image pairs for network training. For the testing, we take 20 PAN/LRMS/GT image pairs with the sizes of ($256 \times 256 \times 1$), ($64 \times 64 \times 8$), ($256 \times 256 \times 8$) on the reduced-resolution evaluation, and 20 PAN/LRMS image pairs with the sizes of ($512 \times 512 \times 1$)/($128 \times 128 \times 8$) thanks to the absence of GT images on the full-resolution assessment.

Metrics. According to the research standard of the pansharpening community, we adopt four quality indexes for the reduced-resolution assessment, including the spectral angle mapper (SAM) [35], ERGAS[36], SCC[37] and Q_{2n}[38]. In terms of the full-resolution evaluation, we use another three metrics, *i.e.,* D_λ, D_s and QNR[39].

Benchmarks. To assess the performance of our approach, we qualitatively and quantitatively compare the proposed method with current state-of-the-art pansharpening methods, including traditional methods and DNN-based techniques. The traditional algorithms include the BT-H [40], BDSD-PC [41], MTF-GLP-FS [42], and CDIF [43] are implemented. Besides, some representative DNN-based models are also compared, such as PNN [6], DiCNN [9], MSDCNN [17], FusionNet [11], and LagNet [14]. Notably, all DNN-based comparison approaches are trained with the same datasets, while the hyperparameter settings comply with the original papers.

Implementation Details. The proposed model is implemented in PyTorch 2.0 and Python 3.10 using a Linux operating system with an NVIDIA RTX3090 GPU. We adopt Adam optimizer with a dynamic learning rate to train the network, where the learning rate is 0.0003 for the first 500 epochs and becomes 0.1 times of the original one for the next 500 epochs.

C ABLATION STUDY

The purpose of ablation research is to determine whether each component of our proposed framework is necessary.Note that all ablation studies are conducted on the WV3 dataset. We first compare the performance of different convolution operations, including the standard convolution, involution operator [22], and our proposed SDAConv. Then, we investigate different strategies to extract global details: *i.e.* using the proposed FGMM and FGMM without the Fourier transform (without FFT). Additionally, we explore the effects of different input sources.

The qualitative comparisons of the ablation studies are displayed in Fig. 10. It is evidently observed that our baseline model (*i.e.* BiMPan) gains the superior visual performance (revealed by the dark blue error map) compared with other configurations, showing the effectiveness of our model design.

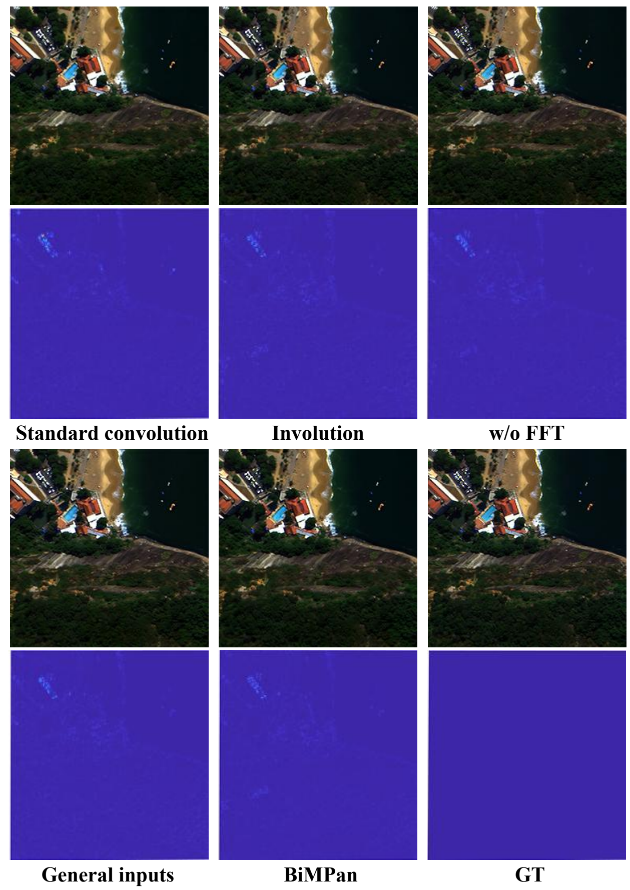

Figure 10: Qualitative comparison of ablation study results on the reduced-resolution sample from WV3 dataset. The first/third row demonstrates the RGB visualization, while the corresponding absolute error maps are presented in the second/fourth row.

D EXPERIMENT ON WV2 DATASET

To confirm our model's generalizability, we use 20 WV2 samples from both full-resolution and reduced-resolution to evaluate all DNN-based models trained on the WV3 dataset. Specifically, we select five representative DNN-based methods for comparison, including PNN [6], DiCNN [9], MSDCNN [17], FusionNet [11], and LagNet [14].

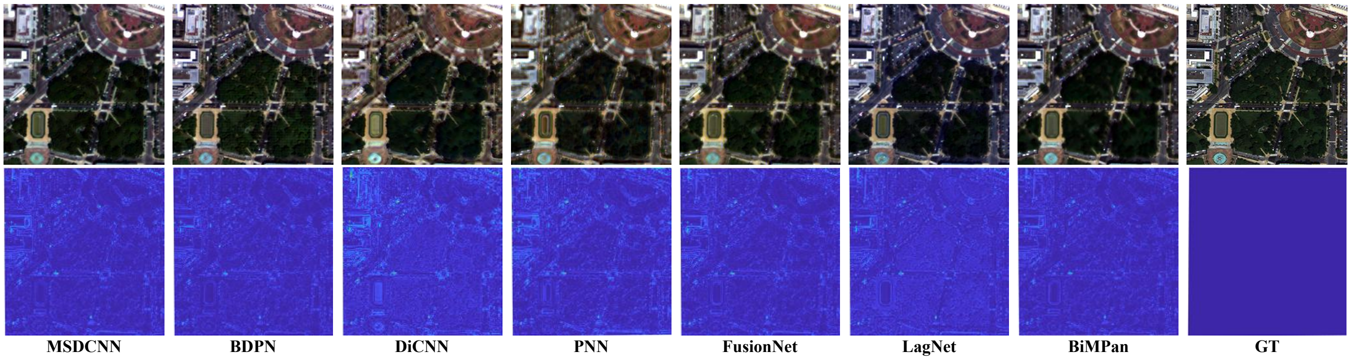

Figure 11: Qualitative comparison on the reduced-resolution sample from WV2 dataset. The first row demonstrates the RGB visualization, while the corresponding absolute error maps are presented in the second row.

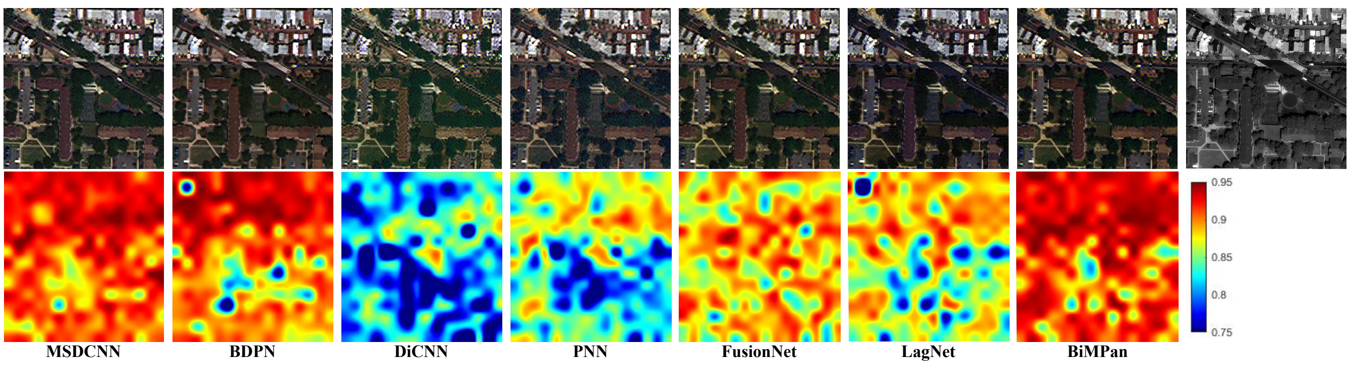

Figure 12: Qualitative comparison on the full-resolution sample from WV2 dataset. The first row presents the RGB visualization, while the second row gives the corresponding QNR maps. The rightmost image is PAN.

The spectral channel of the WV2 sensor and the WV3 sensor are identical, but their spatial resolution is slightly different. WV2 provides eight MS bands and a high-resolution PAN channel. The four standard colors are red, green, blue, and near-infrared 1, while the four new bands that make up these eight bands are coastal, yellow, red edge, and near-infrared 2. Because the PAN images and the MS images are dispersed with pixels that are 0.5 m in size and 2 m in size, respectively, the spatial resolution ratio is equal to 4. 11 bits are utilized in radiometric goal. WV3 and WV2 data share the same channel. However, in contrast to the characteristics of WV2 data, WV3 has spatial resolutions of 1.2 m and 0.3 m.

Therefore, WV2 dataset serves as a perfect choice to test the generalization ability of the networks trained on WV3. As shown in Fig. 11 and Fig. 12, our proposed method demonstrates the favorable visual effect (as illustrated in the dark blue error map and hot QNR map) on WV2 dataset, demonstrating its excellent generalization ability.