Bidomain Modeling Paradigm for Pansharpening

Junming Hou*  Qi Cao*  Ran Ran  Che Liu  Junling Li†  Liang-jian Deng†
School of Information Science and Engineering, Southeast University  Yingcai Honors College, University of Electronic Science and Technology of China  School of Mathematical Sciences, University of Electronic Science and Technology of China
Nanjing, China  Chengdu, China  Nanjing, China
junming_hou@seu.edu.cn  2020080601007@std.uestc.edu.cn  junlingli@seu.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.

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

ACM Reference Format:

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 super-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 representational capacity of pansharpening [8–11]. Although existing DNN-based 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 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.

• 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 ResFFT-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-/upsampling 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
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 $M \in \mathbb{R}^{H \times W \times c}$, we first split it into $c$ single bands. Then, we send each band paired with a PAN image $P$ 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 perform this weight matrix $M_\Omega$ to focus on the local specificity of every band, where $M_\Omega$ denotes the obtained weight matrix.

After that, we perform this weight matrix $M_\Omega$ on a group of candidate kernels $K \in \mathbb{R}^{C_{out} \times k \times k \times C_{in}}$ 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:

$$F_\text{spa}^i, F_\text{spe}^i = SM(\sigma(\text{Cono}(A))_i), \text{CM}(\sigma(\text{Cono}(A))_i),$$

$$W_\Omega = \text{Re}(F_\text{spa}^i \odot F_\text{spe}^i),$$

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

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
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:

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

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:

\[
A(x)(u, v) = \sqrt{R^2(x)(u, v) + I^2(x)(u, v)},
\]

\[
P(x)(u, v) = \frac{I(x)(u, v)}{R(x)(u, v)},
\]

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:

\[
X_R, X_I = \mathcal{F}(X),
\]

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

\[
S_R, S_I = \sigma(DW(X_R)), \sigma(DW(X_I)),
\]

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:

\[
Z_F = \mathcal{F}^{-1}(S_R, S_I).
\]

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

\[
Z_S = \sigma(DW(X)),
\]

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:

\[
U = H_S(Z_F, Z_S),
\]

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 \( 3 \times 3 \) 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:

\[
C_R, C_I = \sigma(MLP(\mathcal{F}(U_R))), \sigma(MLP(\mathcal{F}(U_I)));
\]

\[
C_F = \mathcal{F}^{-1}(C_R, C_I),
\]

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:

\[
\hat{C}_S = \sigma(MLP(U)),
\]

\[
Out = H_c(C_F, \hat{C}_S),
\]

where \( \hat{C}_S \) 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)

<table>
<thead>
<tr>
<th>method</th>
<th>Reduced SAM(± std)</th>
<th>ERGAS(± std)</th>
<th>Q8(± std)</th>
<th>SCC(± std)</th>
<th>Full D3(± std)</th>
<th>QNR(± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDS-D-PC</td>
<td>5.4675±1.7185</td>
<td>4.6549±1.4667</td>
<td>0.8117±0.1063</td>
<td>0.9049±0.0419</td>
<td>0.023±0.0171</td>
<td>0.906±0.0470</td>
</tr>
<tr>
<td>MTF-GLP-FS</td>
<td>5.3233±1.6548</td>
<td>4.6452±1.4441</td>
<td>0.8177±0.1014</td>
<td>0.8984±0.0466</td>
<td>0.035±0.0211</td>
<td>0.904±0.0454</td>
</tr>
<tr>
<td>BT-H</td>
<td>4.8985±1.3028</td>
<td>4.5150±1.3315</td>
<td>0.8182±0.1019</td>
<td>0.9240±0.0243</td>
<td>0.043±0.0232</td>
<td>0.880±0.0340</td>
</tr>
<tr>
<td>CDIF</td>
<td>4.8548±1.4788</td>
<td>4.5029±1.5338</td>
<td>0.8322±0.1032</td>
<td>0.9163±0.0298</td>
<td>0.031±0.0075</td>
<td>0.939±0.0213</td>
</tr>
<tr>
<td>PNN</td>
<td>3.6798±0.7625</td>
<td>2.6819±0.6475</td>
<td>0.8929±0.0923</td>
<td>0.9761±0.0075</td>
<td>0.021±0.0080</td>
<td>0.936±0.0212</td>
</tr>
<tr>
<td>DiCNN</td>
<td>3.5929±0.7623</td>
<td>2.6733±0.6627</td>
<td>0.9004±0.0871</td>
<td>0.9763±0.0072</td>
<td>0.036±0.0111</td>
<td>0.919±0.0258</td>
</tr>
<tr>
<td>MSDCNN</td>
<td>3.7773±0.8032</td>
<td>2.7608±0.6884</td>
<td>0.8900±0.0900</td>
<td>0.9741±0.0076</td>
<td>0.023±0.0091</td>
<td>0.931±0.0271</td>
</tr>
<tr>
<td>BDPN</td>
<td>4.1646±0.8223</td>
<td>3.0335±0.7269</td>
<td>0.8724±0.0979</td>
<td>0.9677±0.0087</td>
<td>0.039±0.0251</td>
<td>0.916±0.0404</td>
</tr>
<tr>
<td>FusionNet</td>
<td>3.3252±0.6978</td>
<td>2.4666±0.6446</td>
<td>0.9048±0.0904</td>
<td>0.9807±0.0069</td>
<td>0.023±0.0090</td>
<td>0.936±0.0137</td>
</tr>
<tr>
<td>LagNet</td>
<td>3.1042±0.5585</td>
<td>2.2999±0.6128</td>
<td>0.9098±0.0907</td>
<td>0.9838±0.0068</td>
<td>0.036±0.0148</td>
<td>0.923±0.0247</td>
</tr>
<tr>
<td>BiMPan(ours)</td>
<td>2.9842±0.6009</td>
<td>2.2569±0.5520</td>
<td>0.9153±0.0865</td>
<td>0.9843±0.0049</td>
<td>0.017±0.0128</td>
<td>0.949±0.0255</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>method</th>
<th>Reduced</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAM(± std) ERGAS(± std) Q4(± std) SCC(± std) D3(± std) D4(± std) QNR(± std)</td>
<td></td>
</tr>
<tr>
<td>BDS-DPC</td>
<td>8.262±0.0497 7.5420±0.8138 0.838±0.1013 0.903±0.0118 0.0345±0.0172 0.1636±0.0483 0.8078±0.0497</td>
<td></td>
</tr>
<tr>
<td>MTF-GLP-FS</td>
<td>8.113±1.9553 7.5102±0.7926 0.829±0.0905 0.898±0.0196 0.0570±0.0137 0.1500±0.0238 0.8017±0.0295</td>
<td></td>
</tr>
<tr>
<td>BT-H</td>
<td>7.194±1.3523 7.4008±0.8738 0.832±0.0880 0.915±0.0152 0.052±0.0141 0.1648±0.0167 0.7912±0.0177</td>
<td></td>
</tr>
<tr>
<td>CDIF</td>
<td>7.296±1.6703 7.1086±0.7077 0.846±0.0918 0.911±0.0139 0.017±0.0137 0.0486±0.0298 0.9351±0.0998</td>
<td></td>
</tr>
<tr>
<td>PNN</td>
<td>5.205±0.9625 4.4722±0.3734 0.918±0.0938 0.971±0.0123 0.056±0.0112 0.0624±0.0239 0.8844±0.0304</td>
<td></td>
</tr>
<tr>
<td>DiCNN</td>
<td>5.379±1.0266 5.1354±0.4876 0.904±0.0942 0.962±0.0133 0.092±0.0143 0.1067±0.0210 0.811±0.0310</td>
<td></td>
</tr>
<tr>
<td>MSDCNN</td>
<td>5.147±0.9342 4.3828±0.3400 0.917±0.0987 0.972±0.0124 0.032±0.0237 0.0667±0.0282 0.9041±0.0466</td>
<td></td>
</tr>
<tr>
<td>BDNP</td>
<td>6.122±1.2106 5.2756±0.6870 0.899±0.0938 0.958±0.0154 0.073±0.0273 0.0492±0.0126 0.8812±0.0336</td>
<td></td>
</tr>
<tr>
<td>FusionNet</td>
<td>4.922±0.9077 4.1594±0.3212 0.925±0.0902 0.975±0.0104 0.058±0.0189 0.052±0.0088 0.8922±0.0219</td>
<td></td>
</tr>
<tr>
<td>LagNet</td>
<td>4.558±0.8155 3.8436±0.4032 0.930±0.0935 0.980±0.0088 0.084±0.0238 0.0676±0.0136 0.8536±0.0178</td>
<td></td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>method</th>
<th>Reduced</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAM(± std) ERGAS(± std) Q8(± std) SCC(± std) D3(± std) D4(± std) QNR(± std)</td>
<td></td>
</tr>
<tr>
<td>PNN</td>
<td>7.118±0.6812 5.6152±0.9431 0.7619±0.0928 0.878±0.0175 0.148±0.0097 0.071±0.0169 0.7869±0.0959</td>
<td></td>
</tr>
<tr>
<td>DiCNN</td>
<td>6.921±0.7898 6.2507±0.5745 0.7205±0.0746 0.855±0.0289 0.141±0.0661 0.1023±0.0195 0.7700±0.0505</td>
<td></td>
</tr>
<tr>
<td>MSDCNN</td>
<td>6.006±0.6377 4.7438±0.4939 0.8241±0.0799 0.897±0.0109 0.058±0.0421 0.029±0.0138 0.9143±0.0516</td>
<td></td>
</tr>
<tr>
<td>BDNP</td>
<td>7.093±0.8630 4.8568±0.5698 0.8235±0.0929 0.903±0.0094 0.111±0.0859 0.032±0.0243 0.8606±0.0979</td>
<td></td>
</tr>
<tr>
<td>FusionNet</td>
<td>6.4257±0.8602 5.1363±0.5151 0.796±0.0737 0.874±0.0134 0.051±0.0092 0.055±0.0146 0.8948±0.0187</td>
<td></td>
</tr>
<tr>
<td>LagNet</td>
<td>6.954±0.4739 5.3262±0.3185 0.8054±0.0837 0.912±0.0101 0.130±0.0856 0.0547±0.0159 0.8229±0.0884</td>
<td></td>
</tr>
</tbody>
</table>

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.
Different Strategies of Global Detail Extraction. We conduct

Table 4: Qualitative comparisons of the different convolution operations.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SAM(± std)</th>
<th>ERGAS(± std)</th>
<th>Q8(± std)</th>
<th>SCC(± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard convolution</td>
<td>3.30±0.68</td>
<td>2.47±0.67</td>
<td>0.91±0.09</td>
<td>0.98±0.01</td>
</tr>
<tr>
<td>Involution</td>
<td>3.04±0.61</td>
<td>2.28±0.58</td>
<td>0.92±0.09</td>
<td>0.98±0.01</td>
</tr>
<tr>
<td>SDAConv</td>
<td>2.98±0.60</td>
<td>2.26±0.55</td>
<td>0.92±0.09</td>
<td>0.98±0.01</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SAM(± std)</th>
<th>ERGAS(± std)</th>
<th>Q8(± std)</th>
<th>SCC(± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o FFT</td>
<td>3.16±0.61</td>
<td>2.46±0.63</td>
<td>0.91±0.09</td>
<td>0.98±0.01</td>
</tr>
<tr>
<td>FGMM</td>
<td>2.98±0.60</td>
<td>2.26±0.55</td>
<td>0.92±0.09</td>
<td>0.98±0.01</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Input sources</th>
<th>SAM(± std)</th>
<th>ERGAS(± std)</th>
<th>Q8(± std)</th>
<th>SCC(± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General inputs</td>
<td>3.21±0.65</td>
<td>2.43±0.63</td>
<td>0.92±0.09</td>
<td>0.98±0.01</td>
</tr>
<tr>
<td>HFC</td>
<td>2.98±0.60</td>
<td>2.26±0.55</td>
<td>0.92±0.09</td>
<td>0.98±0.01</td>
</tr>
</tbody>
</table>

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


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 $\text{SR}$ and the ground truth images $\text{GT}$ during the network training process, which can be represented as follows:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \| \text{SR}_i - \text{GT}_i \|_1,$$

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://liangjian-deng.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 QNR[39]. In terms of the full-resolution evaluation, we use another three metrics, i.e., $D_1$, $D_2$, 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], MTIF-GLF-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 ablations 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.](image)

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].
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.