Spatial-Spectral Dual Back-Projection Network for Pansharpening

Kai Zhang, *Member,* IEEE, Anfei Wang, Feng Zhang, Wenbo Wan, Jiande Sun, Lorenzo Bruzzone, *Fellow,* IEEE

***Abstract*—Deep unfolding networks have obtained satisfactory performance in the pansharpening task owing to their sufficient interpretability. Inspired by the back-projection (BP) mechanism, we propose a BP-driven model, spatial-spectral dual back-project network (S^{2}DBPN), to fuse the low spatial resolution multispectral (LR MS) and the high spatial resolution panchromatic (PAN) images by exploiting the BP in spatial and spectral domains. Specifically, the proposed S^{2}DBPN is made up of a spatial BP network, a spectral BP network, and a reconstruction network. In the spatial BP network, spatial down- and up-projection modules are derived from BP, which is responsible for the projection of the LR MS image into the spatial domain. By analogy with the spatial BP, we reformulate the degradation between high spatial resolution multispectral (HR MS) and PAN images as spectral down- and up-projections. Then, the spectral BP network is constructed for the projection of the PAN image along the channel dimension. Finally, the features from spatial and spectral BP networks are integrated to produce the desired HR MS image through the reconstruction network. Compared to the state-of-the-art methods, extensive experiments on QuickBird, GeoEye-1, and WorldView-2 datasets demonstrate that our S^{2}DBPN produces better HR MS images in terms of qualitative and quantitative evaluation metrics. The code of S^{2}DBPN is released at: https://github.com/RSMagneto/S2DBPN.**

***Index Terms*—Pansharpening, spatial back-projection network, spectral back-projection network, remote sensing.**

I. Introduction

WITH the rapid development of remote sensing imaging techniques, many high-resolution satellites have been launched successfully. More and more remote sensing images are obtained and applied to various fields. However, optical and multi/hyperspectral images often suffer from the limitations of spatial and spectral resolutions [1]-[2]. The physical tradeoff in imaging sensors limits the concurrent increase of spatial and spectral resolutions of remote sensing images [3]-[4]. For example, given a specific sensor, the spatial resolution of the panchromatic (PAN) image is higher than that of the low spatial resolution multispectral (LR MS) image. However, the LR MS image contains abundant spectral information because it is composed of several bands, e.g. 4 bands. By contrast, the PAN image consists of only one channel. In this context, pansharpening techniques are advanced to produce high spatial resolution multispectral (HR MS) image, which integrates the spatial and spectral information present in both PAN and LR MS images.

Over the past three decades, a variety of pansharpening methods have been proposed. They can be classified into four categories [5] including component substitution (CS), multiresolution analysis (MRA), variational optimization (VO), and deep neural network (DNN).

For the methods belonging to the first category, they separate the up-sampled LR MS image spatial and spectral components via some transforms. Then, the spatial component is substituted by the corresponding PAN image. Next, the inverse transform is conducted on the replaced components to synthesize the desired HR MS image. For these methods, it is important to select a proper transform as it significantly affects the quality of the obtained HR MS image. The commonly used transforms are intensity-hue-saturation (IHS) [6]-[7], Gram-Schmidt (GS) transform [8]-[9], principal component analysis (PCA) [10]-[11], and band-dependent spatial detail (BDSD) [12]-[13]. The primary advantage of CS-based methods is the simplicity of implementation. However, spectral distortions often appear in their fusion results owing to differences between PAN and LR MS images in terms of the spectral range.

Different from CS-based methods, the techniques based on MRA only extract the spatial details from the PAN image and then inject them into the up-sampled LR MS image. These methods assume that the missing spatial information in the LR MS image can be obtained from the PAN image. MRA tools, such as wavelet [14]-[15], contourlet [16], and curvelet [17] are extensively explored to improve fusion performance because they can effectively describe high frequencies and spatial details in the images. In addition, some MRA-like pansharpening methods are also developed to extract more reasonable spatial details in the PAN image, such as support value transform (SVT) [18], multiscale nonlocal means filter [19], and support tensor transform [20]. These methods preserve the spectral features in the fused image better because only the spatial details from the PAN image are introduced into the LR MS image.

This work was supported in part by the Natural Science Foundation of China (61901246), the China Postdoctoral Science Foundation Grant (2019TQ0190, 2019M662432), the Scientific Research Leader Studio of Ji'nan (2021GXRC081), and Joint Project for Smart Computing of Shandong Natural Science Foundation (ZR2020LZH015). (*Corresponding author: Kai Zhang.*)

K. Zhang, A. Wang, F. Zhang, W. Wan, and J. Sun are with the School of Information Science and Engineering, Shandong Normal University, Ji'nan 250358, China (e-mail: zhangkainuc@163.com, wanganfei1009@163.com, fengzhangpl@163.com, wanwenbo@sdnu.edu.cn, jiandesun@hotmail.com).

L. Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

In the third category, the fused image is derived from the spatial and spectral degradation models, which are solved by VO. The pansharpening methods in this category regard the LR MS image as the spatial degradation result of the HR MS image. Similarly, the PAN image is the spectral degradation result of the HR MS image. Then the HR MS image is restored by solving the degradation model. But the ill-posedness of the degradation model cannot be ignored. Thus, the priors in source images and the HR MS image are mined to alleviate its ill condition. VO-based methods often adopt sparsity [21]-[22], low rank [23]-[25], and variation prior [26]-[27] to produce a more accurate HR MS image. VO-based methods behave well in terms of the preservation of spatial and spectral features. However, their optimization is generally time-consuming because of the adopted alternative and iterative algorithms.

Following their success on various tasks, DNNs have also been applied to the fusion of LR MS and PAN images [28]. DNN-based pansharpening methods aim to learn a nonlinear mapping between source images and the corresponding HR MS image. For example, Masi *et al.* [29] tried to use a convolution neural network (CNN) to produce the pansharpened MS image. He *et al.* [30] designed a CNN to learn the spatial details missing in the LR MS image. Two CNNs were considered in [31] to extract features from LR MS and PAN images. Then, these features were merged by cascaded adaptive fusion modules. In addition, residual learning was adopted in [32] to model spatial information better. Huang *et al.* [33] combined MRA and residual learning to extract more robust spatial and spectral features from the source images. Lei *et al.* [34] proposed a nonlocal attention residual network to better consider the global similarity in images.

With the advent of new architectures, the generative adversarial network (GAN) is also used for the pansharpening task. For instance, Shao *et al.* [35] utilized a GAN to better preserve the spatial information in the fused image. Meanwhile, a similar framework was also proposed in [36]. Diao *et al.* [37] presented a GAN to fuse LR MS and PAN images without training in advance. In this method, multiscale generators were constructed to enrich the spatial details in the fused image progressively. Li *et al.* [38] employed a cycle-consistent GAN to achieve unsupervised training on unpaired datasets. Furthermore, recent networks based on the transformer are drawing the attention of researchers. Meng *et al.* [39] first applied the vision transformer to the pansharpening task. The transformer with shifted windows (Swin transformer) was then used in [40]. Sun *et al.* [41] also established a regression network based on the Swin transformer to reconstruct the HR MS image. CNN and transformer were incorporated in [42] to learn the local and global information in source images simultaneously.

Besides, deep unfolding networks have also been explored owing to their interpretability and effectiveness [43]. In these methods, the optimization of degradation models is unfolded as a DNN to capture the model prior and image prior. Xu *et al.* [44] developed a deep gradient projection network (DGPNN) that was composed of a series of cascaded MS and PAN blocks. Yang *et al.* [45] derived a deep conditional unfolding network from the degradation model and nonlocal similar prior. Cao *et al.* [46] presented an optimization algorithm based on convolutional sparse coding, which was unfolded as a DNN with interpretable structures. Dian *et al.* [47]-[48] combined DNNs with the optimization of degradation models, which can learn the spatial and spectral priors in images efficiently. Tian *et al.* [49] exploited the similarity between the fused image and the PAN image in the gradient domain and constructed a variational pansharpening network from the degradation model regularized by the similarity prior.

Recently, back projection (BP) has been employed to facilitate the construction of deep unfolding networks. In BP, residual images are first iteratively computed as the reconstruction error between the target image and its counterparts. Then, the residual images are back-projected into the target image to improve its spatial details [50]. Following the paradigm, BP-driven DNNs are built. For example, Haris *et al.* [51] proposed a deep back-projection network (DBPN) for image super-resolution which was made up of stacked up- and down-sampling units. Subsequently, DBPN was further promoted for the super-resolution of videos [52]. In [53], hybrid residual features were introduced into DBPN to improve its compactness. Liu *et al.* [54] designed a weight module to integrate the features at different levels together, by which the spatial information in these features was sufficiently exploited.

Compared to other deep unfolding networks, BP-driven ones do not involve complicated optimization strategies and are simple in principle although the degradation model is also used in these methods. Therefore, considering the state-of-the-art performance and simplicity of BP-driven DNNs, we use this formulation to generate the fused image. However, only the spatial degradation between HR and LR images is considered in existing BP-driven DNNs, which cannot deal with the degradation in the spectral domain. Apart from the spatial degradation between LR MS and HR MS images, the spectral degradation between PAN and HR MS images is crucial in the pansharpening task. Thus, the aforementioned BP-driven DNNs cannot be directly applied to the fusion of LR MS and PAN images due to the dual degradation models in spatial and spectral domains.

To overcome the above-mentioned limitations, in this paper we propose a spatial-spectral dual back-project network (S^{2}DBPN) for pansharpening. In the proposed S^{2}DBPN, spatial and spectral BP networks are designed to project the LR MS and PAN images, respectively. In the spatial BP network, spatial down- and up-projection modules are constructed according to the BP between LR MS and HR MS images. To enhance the spatial details in the features of the LR MS image, spatial-aware blocks are introduced into these modules. Similar to the projections in the spatial BP network, spectral down- and up-projection modules are integrated into the spectral BP network, which projects the feature of the PAN image along the channel dimension. In the spectral BP network, the multi-head mechanism is considered to increase the number of channels in the feature space, which makes full use of the information from different subspaces. Finally, all features from spatial and spectral BP networks are concatenated and fed into the reconstruction network for the generation of the HR MS image. To the best of our knowledge, this paper is the first to embed the spatial BP and spectral BP into DNN for pansharpening. Compared to existing DNN-based methods, the proposed S^{2}DBPN is derived from BP, whose interpretability is ensured. The blocks in spatial and spectral BP networks correspond to the down- and up-sampling operations in spatial and spectral BP, respectively. Moreover, the experiments on reduced- and full-scale QuickBird, GeoEye-1, and WorldView-2 datasets demonstrate the effectiveness of the proposed S^{2}DBPN.

The main contributions of the paper are summarized as follows:

1) We propose a BP-driven model, S^{2}DBPN that adopts the formulation of BP, to produce the desired HR MS images from LR MS and PAN images. Compared to other deep unfolding networks, the proposed S^{2}DBPN does not require complicated optimization algorithms and can be trained effectively.

2) We design a spatial BP network to implement the BP between HR MS and LR MS images in the spatial domain. Spatial network modules are constructed to learn the spatial down- and up-sampling projections in BP adaptively.

3) We design a spectral BP network to enhance the feature of the PAN image along the channel dimension by spectral down- and up-projection modules. The network is derived from the spectral degradation between HR MS and PAN images.

The rest of the paper is organized as follows. In Section II the BP is briefly introduced. Section III presents the proposed S^{2}DBPN in detail. Section IV demonstrates extensive results on datasets from different satellites. Finally, conclusions are given in Section V.

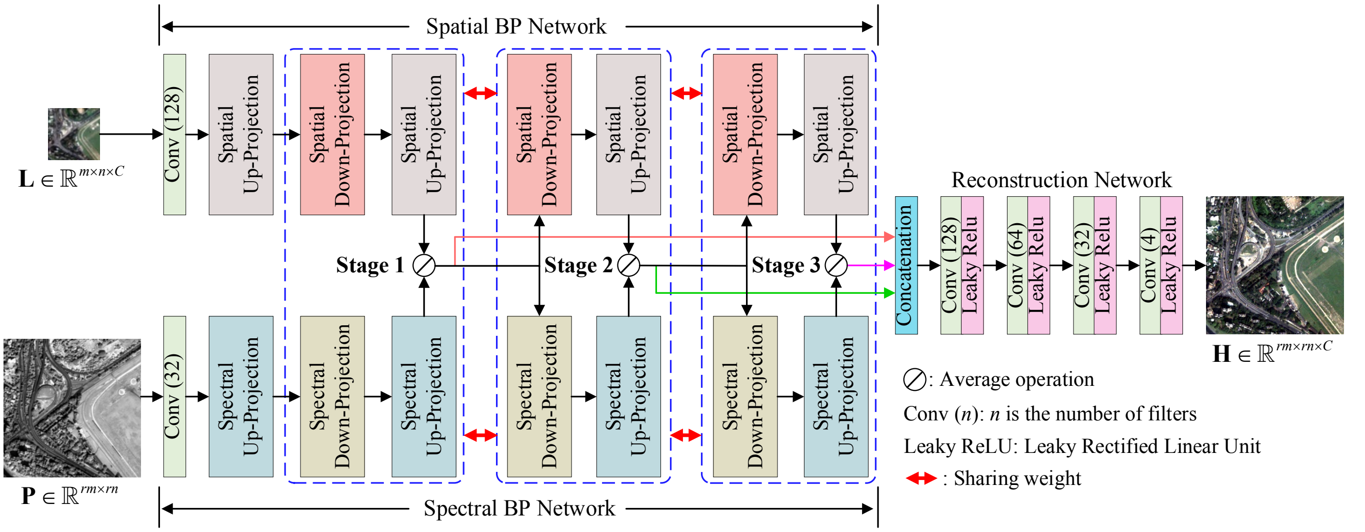


Fig. 1. Architecture of the proposed S^{2}DBPN.

II. Back Projection

As a typical image refinement technique, BP improves the consistency between LR and HR images in terms of the degradation model by back-projecting the reconstruction errors into the HR image. The formulation of BP can be represented as:

$$\begin{aligned} I_i^l &= \left(I_i^h * g\right)\downarrow_d \\ E_i^l &= I^l - I_i^l \\ E_i^h &= \left(E_i^l \uparrow_d\right) * p \\ I_{i+1}^h &= I_i^h + E_i^h \end{aligned} \tag{1}$$

where the spatial up- and down-sampling operators are $\uparrow_d$ and $\downarrow_d$. d is the sampling ratio. Because spatial up- and down-sampling are involved, the spatial dimensions of images are changed in (1). Thus, we superscripts, l and h, to represent LR and HR, respectively. I^l is the observed LR image and I_i^l is the LR version after the spatial degradation of the HR image I_i^h in the ith iteration. E_i^l is the difference between I^l and I_i^l. E_i^h is the back-projected HR version of E_i^l. g and p are the blur kernel and the back-projection kernel, respectively. The convolution operation is denoted by $*$.

In the ith iteration of BP, the HR image I_i^h is first spatially degraded as I_i^l. Then, the reconstruction error E_i^l is obtained from the difference between I^l and I_i^l . Finally, the up-sampled reconstruction error E_i^h is injected into I_i^h to refine the HR image. One can see that the quality improvement of the HR image is dependent on the handcrafted g and p in (1). To further enhance the HR images, BP-driven DNNs aim to learn these kernels adaptively in the feature space considering the powerful mapping ability of DNNs [51]-[54].

III. Spatial-Spectral Dual Back-Projection Network

This section presents the technical details of the proposed S^{2}DBPN, including the spatial BP and the spectral BP. Then, the desired HR MS image is reconstructed by integrating all features from different stages.

A. Overall Architecture

Fig. 1 shows the architecture of the proposed S^{2}DBPN. It is composed of a spatial BP network, a spectral BP network, and a reconstruction network. In spatial and spectral BP networks, the LR MS image $\mathbf{L} \in \mathbb{R}^{m\times n\times C}$ and the PAN image $\mathbf{P} \in \mathbb{R}^{rm\times rn}$ are fed into the corresponding convolution layers to obtain a feature embedding, respectively. In the convolution layers, the size of the filter is 3×3 . m and n are the width and height of the LR MS image. C denotes the number of bands in the LR MS image. The spatial size of the PAN image is $rm\times rn$. r is the spatial ratio between LR MS and PAN images and is typically set as 4. For the PAN image, the extracted features are sent into the spectral up- and down-projection modules to achieve BP in the spectral domain. We employ spatial up- and down-projection modules to map the LR MS image for spatial BP. Then, an average operation is utilized in each stage to fuse the feature maps from spectral and spatial up-projection modules. By BP along spatial and spectral dimensions, the residuals are iteratively computed to enhance the spatial and spectral details of the fused images in feature space. Learning the up- and down-projections in BP can efficiently capture the relationships between the HR MS image and source images, which further improves the consistency among them. In the proposed S^{2}DBPN, spatial and spectral BP is achieved by successive three stages. The same structures are used in these stages, which also share the same model parameters. Finally, all features in different stages are concatenated to generate the HR MS image $\mathbf{H}\in\mathbb{R}^{rm\times rn\times C}$ via the reconstruction network, in which the size of filters in all convolution layers is 3×3 . The following parts will present the details of the proposed S^{2}DBPN. Table I summarizes some important symbols used in the following sections.

TABLE I. DEFINITIONS OF SYMBOLS.

Symbols	Definitions
$\mathcal{D}(\cdot)$	Spatial degradation operator
$\mathcal{S}(\cdot)$	Spectral degradation operator
$\mathcal{V}_{down}^{i}(\cdot)$, $\mathcal{T}_{down}^{i}(\cdot)$	Down-sampling operators in spatial down-projection module
$\mathcal{V}_{up}^{i}(\cdot)$	Up-sampling operator in spatial down-projection module
$\mathcal{U}_{up}^{i}(\cdot)$, $\mathcal{S}_{up}^{i}(\cdot)$	Up-sampling operators in spatial up-projection module
$\mathcal{U}_{down}^{i}(\cdot)$	Down-sampling operator in spatial up-projection module
$\mathcal{Q}_{down}^{i}(\cdot)$, $\mathcal{J}_{down}^{i}(\cdot)$	Down-sampling operators in spectral down-projection module
$\mathcal{Q}_{up}^{i}$	Up-sampling operator in spectral down-projection module
$\mathcal{P}_{up}^{i}(\cdot)$, $\mathcal{I}_{up}^{i}(\cdot)$	Up-sampling operators in spectral up-projection module
$\mathcal{P}_{down}^{i}$	Down-sampling operator in spectral up-projection module

B. Spatial BP Network

In the pansharpening task, the degradation relationship between LR MS and HR MS images is similar to that between LR and HR images in Section II. The spatial degradation is modeled as:

$$\mathbf{L}=\mathcal{D}(\mathbf{H})+e_1 \tag{2}$$

where $\mathcal{D}(\mathbf{H})$ denotes the spatial degradation of the HR MS image including blurring and down-sampling operations. Typically, the down-sampling ratio is 4. e_1 is the Gaussian noise.

Although the spatial information of the LR MS image can be improved by the BP in (1), simple up- and down-sampling operators cannot describe the complicated mapping between them, which limits the quality of the HR MS image. To better reconstruct the HR MS image, DNNs are introduced as the up- and down-sampling operators to obtain deep features for the generation of the HR MS image. According to the formulation of BP in (1), we utilize three stages consisting of spatial down- and up-projection modules to enrich the spatial information in features. In each stage, the HR feature from the previous spatial up-projection module is projected as the LR version, which is enhanced to obtain the refined HR feature through the following spatial up-projection module.

1) Spatial Down-Projection: In the spatial down-projection module of the ith stage, the HR feature M^{i-1} from the previous spatial up-projection module is down-sampled to obtain the LR version L^i and the process is defined as:

$$\begin{aligned}
&M_{down}^{i}=\mathcal{V}_{down}^{i}\left(M^{i-1}\right)\\
&M_{up}^{i}=\mathcal{V}_{up}^{i}\left(M_{down}^{i}\right)\\
&F_{up}^{i}=M_{up}^{i}-M^{i-1}\\
&F_{down}^{i}=\mathcal{T}_{down}^{i}\left(F_{up}^{i}\right)\\
&L^{i}=M_{down}^{i}+F_{down}^{i}
\end{aligned} \tag{3}$$

where $\mathcal{V}_{down}^{i}$ is a down-sampling operator with a ratio of r and the LR feature M_{down}^{i} is up-sampled by the operator $\mathcal{V}_{up}^{i}$ to produce M_{up}^{i} . Then, the error F_{up}^{i} between M_{up}^{i} and M^{i-1} is down-sampled by the operator $\mathcal{T}_{down}^{i}$ and back-projected to M_{down}^{i} .

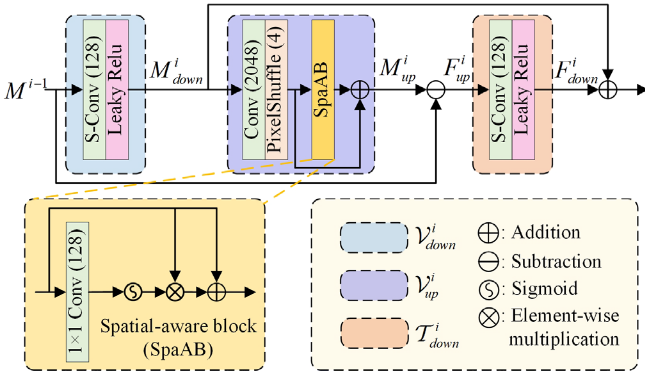


Fig. 2. Architecture of the spatial down-projection module.

According to the formulation in (3), we construct the spatial down-projection module illustrated in Fig. 2. In Fig.2, the down- and up-sampling operators in (3) are made up of different network blocks. In $\mathcal{V}_{down}^{i}$, a strided convolution layer (S-Conv) is introduced to down-sampling the HR feature M^{i-1} . In S-Conv, 128 filters with a size of 8×8 are equipped and the stride is 4. Then, the width and height of M_{down}^{i} shrink to a quarter of those of M^{i-1} . In $\mathcal{V}_{up}^{i}$, the pixel shuffle technique is adopted to resize the LR feature M_{down}^{i} to the dimensions of the HR feature. PixelShuffle (4) means that the features are up-sampled with a ratio of 4. Moreover, a spatial-aware block

(SpaAB) is designed to further enhance the spatial information in the HR feature. In SpaAB, the spatial information in features is highlighted by 128 filters with the size of 1×1 and then combined the attention map with itself. $\mathcal{T}_{down}^{i}$ adopts the same structure as that of $\mathcal{V}_{down}^{i}$. Finally, the down-sampled feature F_{down}^{i} is injected into M_{down}^{i} for down-projection.

2) Spatial Up-Projection: Following the BP paradigm in (1), the spatial up-projection between LR MS and HR MS images is designed as:

$$\begin{aligned} L_{up}^{i} &= \mathcal{U}_{up}^{i}\left(L^{i}\right) \\ L_{down}^{i} &= \mathcal{U}_{down}^{i}\left(L_{up}^{i}\right) \\ E_{down}^{i} &= L_{down}^{i} - L^{i} \\ E_{up}^{i} &= \mathcal{S}_{up}^{i}\left(E_{down}^{i}\right) \\ M^{i} &= L_{up}^{i} + E_{up}^{i} \end{aligned} \tag{4}$$

where L^{i} is the feature from the spatial down-projection module in the *i*th stage. Through the operator $\mathcal{U}_{up}^{i}$, L^{i} is up-sampled with a ratio of *r* to produce the HR feature L_{up}^{i}. Then, the degraded version L_{down}^{i} is obtained by $\mathcal{U}_{down}^{i}$. In the final step, the error E_{down}^{i} between L_{down}^{i} and L^{i-1} is up-sampled by $\mathcal{S}_{up}^{i}$ and back-projected into L_{up}^{i}.

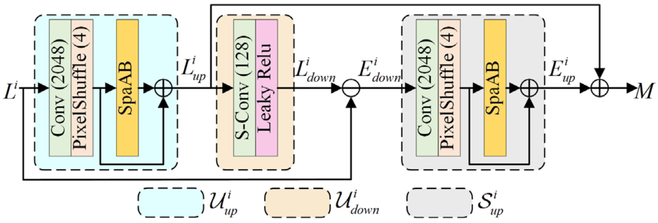


Fig. 3. Architecture of the spatial up-projection module.

Following the projection in (4), we derive the spatial up-projection module shown in Fig. 3, where blocks similar to those in Fig. 2 are assembled for the up-projection of L^{i}. $\mathcal{U}_{up}^{i}$ and $\mathcal{S}_{up}^{i}$ both contain the pixel shuffle layer and a SpaAB. The down-sampling in $\mathcal{U}_{down}^{i}$ is achieved by the S-Conv layer. Through the spatial up-projection module, the error is back-projected into the up-sampled LR feature to further improve the spatial information.

C. Spectral BP Network

Different from the super-resolution task, dual source images, i.e. LR MS and PAN images, are involved in the pansharpening task. Besides the spatial degradation between HR MS and LR MS images in (2), the PAN image is generally viewed as the degradation result of the HR MS image in the spectral domain, which is written as:

$$\mathbf{P}=\mathcal{S}\left(\mathbf{H}\right)+e_{2} \tag{5}$$

where $\mathcal{S}$ stands for the spectral degradation operator and the Gaussian noise is denoted by e_2.

Inspired by the BP in the spatial domain, we believe that the HR MS image can be restored from the PAN image by the BP in the spectral domain. For MS images, the spectral information is embedded into the interdependency and correlation among their bands. In the spatial BP network, the spatial down- and up-projections are implemented on the spatial dimensions of the feature map in the feature space. By analogy with the formulations in (3) and (4), we propose a spectral BP network to project the feature map along the channel dimension, which captures the correlations among channels in the feature map. In this way, the BP in the spectral domain is achieved. Then, three stages containing spectral up- and down-projection modules are introduced into the spectral BP network, whose structures are detailed in the following part.

1) Spectral Down-Projection: In the spectral down-projection of the *i*th stage, the output of the (*i*-1)th stage N^{i-1} is condensed to produce a degraded version along the channel dimension. Similar to (3), the spectral down-projection is expressed as:

$$\begin{aligned} N_{down}^{i} &= \mathcal{Q}_{down}^{i}\left(N^{i-1}\right) \\ N_{up}^{i} &= \mathcal{Q}_{up}^{i}\left(N_{down}^{i}\right) \\ Y_{up}^{i} &= N_{up}^{i} - N^{i-1} \\ Y_{down}^{i} &= \mathcal{J}_{down}^{i}\left(Y_{up}^{i}\right) \\ P^{i} &= N_{down}^{i} + Y_{down}^{i} \end{aligned} \tag{6}$$

where $\mathcal{Q}_{down}^{i}$ and $\mathcal{Q}_{up}^{i}$ are the spectral down- and up-sampling operators, respectively. The dimension of the channels of N^{i-1} is reduced by $\mathcal{Q}_{down}^{i}$ to generate the spectrally degraded feature N_{down}^{i}. Then, N_{down}^{i} is up-sampled by $\mathcal{Q}_{up}^{i}$ along the channel dimension. Next, the channels of the error Y_{up}^{i} between N_{up}^{i} and N^{i-1} is adjusted by the spectral down-sampling operator $\mathcal{J}_{down}^{i}$. Y_{down}^{i} is finally combined with N_{down}^{i}.

By exploiting (6), we design a spectral down-projection module to implement the above operations as shown in Fig. 4(a). In the spectral down-projection module, the feature map with a size of $rm \times rn \times B$ is first reshaped as $N^{i-1} \in \mathbb{R}^{r^2 mn \times B}$, in which each column contains the values on the same spatial position of all channels. Here, B is specifically set as 128. Considering the projection along the channel dimension, each column in N^{i-1} is down-sampled as a vector with the length of b by the operator $\mathcal{Q}_{down}^{i}$, which is composed of a 1D convolution layer and a Leaky ReLU as shown in Fig. 4(b). Here, the 1D convolution layer is introduced to model the correlations among channels in feature maps. Because the spectral degradation of HR MS images is achieved along the spectral dimension, the 1D convolution layer is also analogically implemented on the channels of feature maps. The size of N_{down}^{i} is $rm \times rn \times b$ and b is set as 32. Then, N_{down}^{i} is up-sampled along the channel dimension by $\mathcal{Q}_{up}^{i}$ to obtain $N_{up}^{i} \in \mathbb{R}^{r^2 mn \times B}$. In $\mathcal{Q}_{up}^{i}$, all columns in N_{down}^{i} are fed into a

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3266799

spectral up-sampling block (SpeUB) containing a 1D convolution layer to capture the correlation among channels. Because the dimension of each column is increased from b to B, we extend the SpeUB to three heads to integrate the information from different subspaces. The structure of $\mathcal{Q}_{up}^{i}$ is shown in Fig. 4(c). With the introduction of multi-head SpeUB, the up-sampling along the channel dimension is more stabilized. $\mathcal{J}_{down}^{i}$ shares the same structure as that of the $\mathcal{Q}_{down}^{i}$. Finally, the down-sampled error Y_{down}^{i} is added to N_{down}^{i} and the down-projected feature is reshaped to the size of $rm \times rn \times b$.

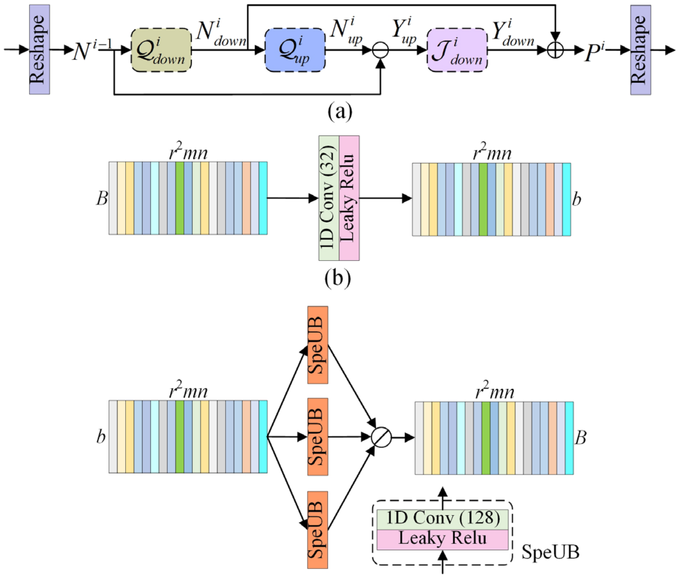


Fig. 4. Architectures of (a) the spectral down-projection module, (b) $\mathcal{Q}_{down}^{i}$, and (c) $\mathcal{Q}_{up}^{i}$.

2) Spectral Up-Projection: The spectral up-projection aims to promote the spectral information in the feature maps by increasing the number of channels. Similar to the spatial up-projection in (4), the spectral up-projection is formulated as:

$$\begin{aligned} P_{up}^{i} &= \mathcal{P}_{up}^{i}\left(P^{i}\right) \\ P_{down}^{i} &= \mathcal{P}_{down}^{i}\left(P_{up}^{i}\right) \\ X_{down}^{i} &= P_{down}^{i} - P^{i} \\ X_{up}^{i} &= \mathcal{I}_{up}^{i}\left(X_{down}^{i}\right) \\ N^{i} &= P_{up}^{i} + X_{up}^{i} \end{aligned} \tag{7}$$

where P^{i} is the output of the spectral down-projection module in the ith stage. The spectral up-sampling operator $\mathcal{P}_{up}^{i}$ first increases the number of channels in P^{i}. Then, the dimension of channels in P_{up}^{i} is reduced by the spectral down-sampling operator $\mathcal{P}_{down}^{i}$. Finally, we improve the error between P_{down}^{i} and P^{i-1} in terms of the channel dimension and add the spectrally up-sampled error X_{up}^{i} into P_{up}^{i}.

Based on the process in (7), we build a spectral up-projection module, as illustrated in Fig. 5. The spectral up- and down-sampling blocks adopt structures similar to those in Fig. 4 to change the channels of the input features. Specifically, the structures of $\mathcal{P}_{up}^{i}$ and $\mathcal{I}_{up}^{i}$ are the same as that of $\mathcal{Q}_{up}^{i}$. $\mathcal{P}_{down}^{i}$ and $\mathcal{Q}_{down}^{i}$ share the same structure. Through the spectral up-projection module, the error is back-projected into the feature with more channels.

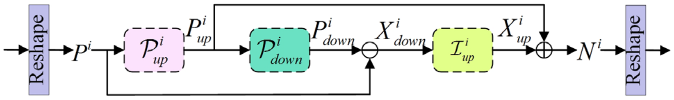


Fig. 5. Architectures of the spectral up-projection module.

D. Reconstruction and Optimization

In the proposed S^{2}DBPN, the three stages shown in Fig. 1 are introduced to complete the BP between the source images and the HR MS image in the feature space. For each stage, their outputs are averaged by (8) to fuse the features from different BP networks:

$$Z^{i} = \frac{Z_{L}^{i} + Z_{P}^{i}}{2} \tag{8}$$

where Z_{L}^{i} and Z_{P}^{i} are the outputs of the spatial and spectral up-projection modules in the ith stage, respectively. Then, Z^{i} is fed into the spatial and spectral down-projection modules in the next stage. Finally, Z^{1}, Z^{2}, and Z^{3} are concatenated to reconstruct the fused image.

Finally, the following loss is minimized to obtain the desired model:

$$\mathcal{L} = \left\| \mathbf{G} - \mathbf{H} \right\|_{F}^{2} \tag{9}$$

where $\mathbf{G}$ is the reference image, i.e. the ground truth. The proposed S^{2}DBPN is trained by the PyTorch framework on a server with Intel® Core™ i7-9700 processor, 3.0 GHz, NVIDIA 2080Ti GPU, and 11-GB memory. In specific, we use Adam as an optimizer, in which the learning rate is 0.00003. Considering the tradeoff between the GPU memory and the computational complexity, the batch size is set as 2. When the optimization reaches 500 epochs, the training is completed.

IV. Experiments

In this section, the effectiveness of the proposed method is demonstrated by the reduced- and full-scale experiments on QuickBird, GeoEye-1, and WorldView-2 datasets. Besides, ablation studies investigate the performance of several variants on the fusion results.

A. Experimental Settings

1) Datasets: Experiments are conducted on two datasets from QuickBird, GeoEye-1, and WorldView-2 satellites. The GeoEye-1 dataset was captured in Hobart, Australia, on February 24, 2009, and contains rural and urban areas. The spatial resolutions of LR MS and PAN images in the dataset are 2.0m and 0.5m, respectively. The QuickBird satellite dataset was collected in the urban area in Sundarbans, India, on November 21, 2002. It includes 2.8m LR MS and 0.7m PAN images. For the WorldView-2 satellite dataset, the spatial resolutions of LR MS and PAN images are 2.0m and 0.5m, respectively. This dataset was taken from the area of Washington DC, USA, on September 26, 2016. To train the

DNN-based methods, the original LR MS and PAN images are down-sampled to the reduced scale to compose the training data according to Wald's protocol [4]. Then, the original MS images are employed as the ground truth. Through the same down-sampling strategy, a set of reduced-scale LR MS and PAN images is generated for reference-based evaluations. Full-scale datasets are also constructed for no-reference evaluations. Table II provides the numbers of training, validation, and test images from different satellites. In these datasets, the sizes of LR MS and PAN images are 64×64 and 256×256, respectively.

2) Compared Methods: The proposed S²DBPN is compared with four traditional methods and four DNN-based methods. The traditional methods are AWLP [15], BDSD [12], GS [4], and MTF-GLP [4]. The DNN-based ones are PNN [29], GPPNN [44], PanNet [55], PSGAN [36], DMDN [56], and Fusion-Net [57]. All DNN-based methods are trained on the same server as that of the proposed S²DBPN.

TABLE II. DETAILS ABOUT TRAINING, VALIDATION, AND TEST DATA.

Satellite	**#Training Pairs**	**#Validation Pairs**	**#Test Pairs**	
			Reduced scale	Full scale
QuickBird	900	20	30	30
GeoEye-1	924	25	30	30
WorldView-2	990	30	30	30

3) Metrics: The fusion results of reduced-scale datasets are evaluated by reference-based metrics, including Q4 [58], root mean square error (RMSE), spectral angle mapper (SAM) [59], universal image quality index (UIQI) [60], and *relative dimensionless global error in synthesis* (ERGAS) [61]. Q4 and UIQI are in a range from 0 to 1 and their optimal value is 1. RMSE, SAM, and ERGAS tend to be 0 for better results. For the evaluation of the full-scale fusion results, D_λ, D_S, and quality no reference (QNR) [62] are considered. D_λ and D_S close to 0 mean that the fused image is better. The best value of QNR is 1.

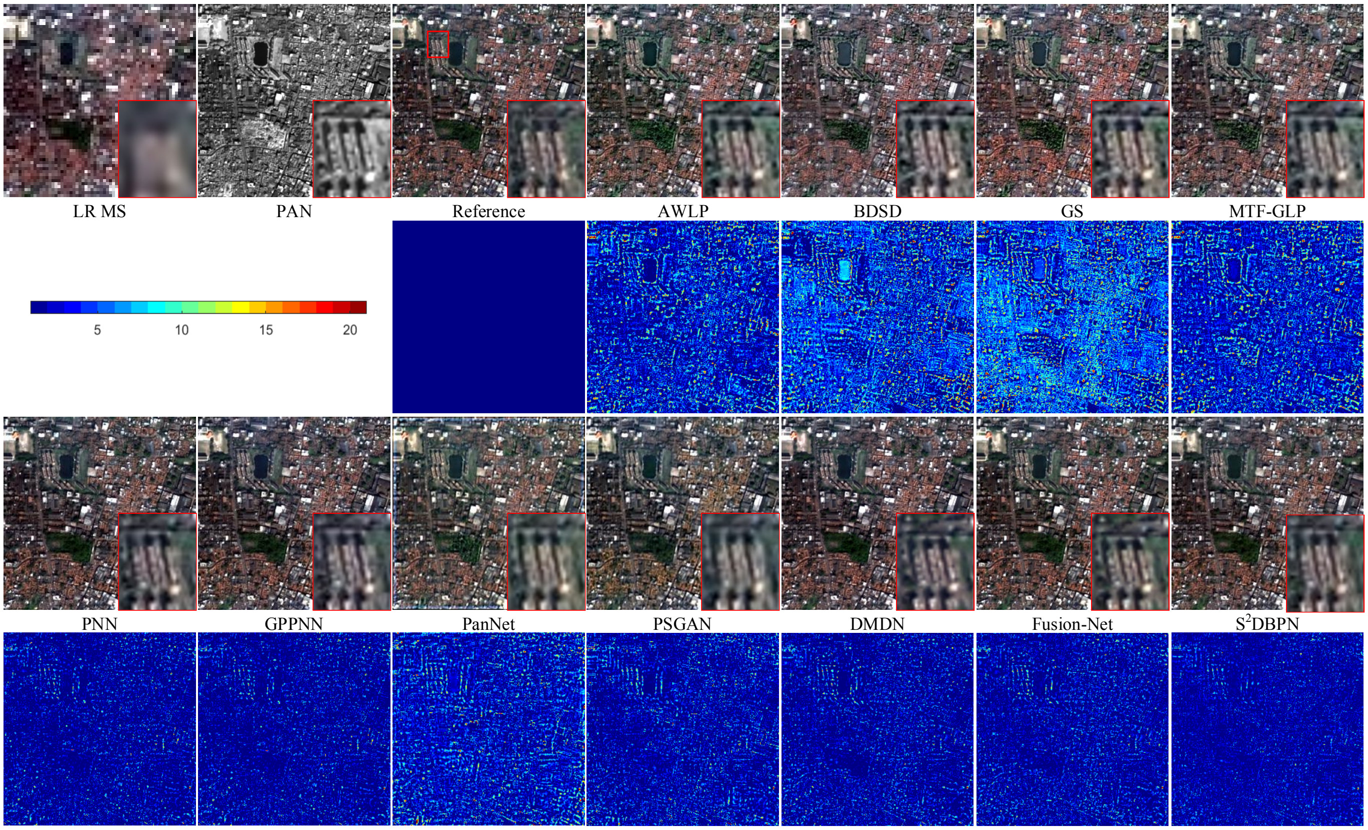


Fig. 6. Qualitative comparison of the fused results of all methods on the reduced-scale QuickBird dataset.

TABLE III. QUANTITATIVE COMPARISON ON THE REDUCED-SCALE DATASET FROM THE QUICKBIRD SATELLITE.

Metric	**AWLP**	**BDSD**	**GS**	**MTF-GLP**	**PNN**	**GPPNN**	**PanNet**	**PSGAN**	**DMDN**	**Fusion-Net**	**S²DBPN**
Q4	0.9198	0.8912	0.8629	0.9161	0.9436	0.9453	0.9038	0.9245	0.9461	0.9420	**0.9477**
RMSE	15.6137	20.4218	19.4334	15.9483	10.9141	10.0751	18.3154	12.9960	10.4769	10.0201	**9.9882**
SAM	2.3588	3.4110	2.9604	2.5355	1.7088	1.5755	2.5797	2.0874	1.6283	1.5797	**1.5658**
UIQI	0.9132	0.9002	0.8470	0.9128	0.9610	**0.9672**	0.8842	0.9511	0.9633	0.9630	**0.9672**
ERGAS	0.8995	1.2033	1.1288	0.9205	0.6280	0.5799	1.0081	0.7494	0.6016	0.5867	**0.5774**

B. *Experiments on Reduced-Scale Datasets*

In this part, the reduced-scale experiments are implemented on the QuickBird, GeoEye-1, and WorldView-2 datasets. Fig. 6 shows the fusion results on the QuickBird dataset. Some visual differences are demonstrated by a local area highlighted by a red rectangle, which is magnified and put in the lower right corner of the fusion result. Moreover, Fig. 6 also shows the absolute difference maps between the fused images and the reference image for the comparison of the reconstruction performance. From Fig. 6, we can see that the spatial details in the fused images of traditional methods are over-enhanced when compared to the reference image, especially in the vegetation regions. For example, the textures of vegetation regions in the AWLP result are excessively sharp. Besides, some spectral distortions also can be seen in the results of BDSD, GS, and MTF-GLP. The absolute difference maps of traditional methods also reflect that their reconstruction errors are large in texture and edge regions. For the PanNet result, one can find some spatial effects similar to those in the results of traditional methods. The reason for this may be that the residual learned by PanNet is over-injected into the up-sampled LR MS image. The building areas in the PSGAN result suffer from spectral distortions, which may be caused by insufficient adversarial learning. Large reconstruction errors also arise in the absolute difference maps of PanNet and PSGAN. PNN, GPPNN, and our S^{2}DBPN have better performance compared to other methods. However, the reconstruction errors of PNN and GPPNN are larger than those in the S^{2}DBPN result. Therefore, the proposed S^{2}DBPN can produce better fusion results.

Table III gives the quantitative metrics associated with the fusion results in Fig. 6, where the average values of each metric on 30 LR MS and PAN image pairs are calculated. The best values in Table III are marked in bold. The best values in Table III illustrate that the proposed S^{2}DBPN behaves better in terms of spatial and spectral metrics.

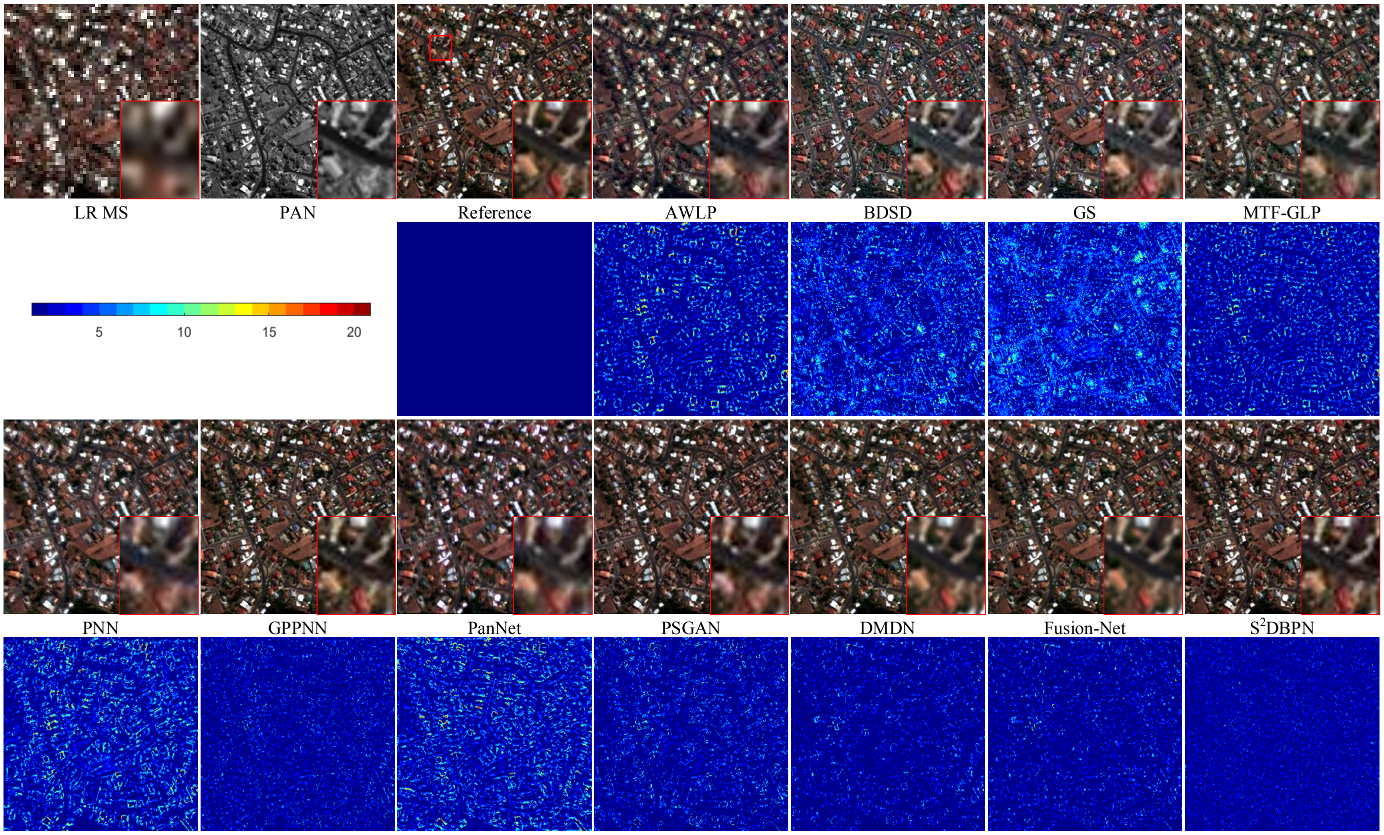


Fig. 7. Qualitative comparison of the fused results of all methods on the reduced-scale GeoEye-1 dataset.

TABLE IV. QUANTITATIVE COMPARISON ON THE REDUCED-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	AWLP	BDSD	GS	MTF-GLP	PNN	GPPNN	PanNet	PSGAN	DMDN	Fusion-Net	S^{2}DBPN
Q4	0.7987	0.7854	0.7644	0.8075	0.7946	0.8230	0.7728	0.8104	0.8107	0.8145	**0.8351**
RMSE	25.5116	29.2840	27.3993	24.5394	27.3163	15.3090	28.5935	22.1315	19.2638	18.7384	**11.1062**
SAM	4.9535	5.7013	4.9287	4.7487	4.7835	2.8742	4.5709	4.2201	3.8100	3.6338	**2.1052**
UIQI	0.9502	0.9419	0.9284	0.9535	0.9383	0.9816	0.9331	0.9610	0.9702	0.9707	**0.9903**
ERGAS	1.5783	1.8290	1.7167	1.5378	1.7010	0.9671	1.7673	1.3789	1.2113	1.1819	**0.7045**

Fig. 7 shows the fusion results of all methods on the reduced-scale GeoEye-1 dataset. Magnified regions and absolute difference maps are also shown in Fig. 7 for further perception. In the AWLP result, some spatial details are lost. For instance, there are some spatial blurring effects in the magnified region of the AWLP result. From the enlarged areas in the BDSD and GS results, we can find some spatial artifacts, which may result from the misestimated gains in the two methods. In the absolute difference maps of AWLP, BDSD, and GS, the large errors are mainly concentrated on the building areas, which validates the loss of spatial details in their fused images. For the MTF-GLP result, the color of the ground is not consistent with that of the reference image. Besides, blurring effects can be observed from PNN, PanNet, and PSGAN results. In the magnified areas of these results, the details of buildings are smoothed. Their corresponding error maps also show that the edges of buildings in the fused images are not preserved well. GPPNN and our S^{2}DBPN have similar performance in terms of visual analysis. But the larger errors in the difference maps of GPPNN imply that the proposed S^{2}DBPN can reconstruct the fused image better.

Table IV reports the values of all reference-based metrics on the reduced-scale GeoEye-1 dataset, where the values of each metric are averaged on 30 LR MS and PAN image pairs. Then, the best values are labeled in bold. The proposed method produces the best values in terms of all metrics. So, the proposed S^{2}DBPN is better than these state-of-the-art methods.

Fig. 8 provides the visual results of all methods on the reduced-scale WorldView-2 dataset. For the results of traditional methods, it can be found that the GS result suffers from obvious spectral distortions in the building areas. For DNN-based methods, PanNet produces large differences in terms of spectral information when compared to the reference image. The error maps in Fig. 8 also demonstrate that GS and PanNet cannot reconstruct the fused images well. Compared to other methods, the reconstruction errors of the proposed method are smaller than those of other methods.

Table V also records the average results of all metrics on the WorldView-2 dataset. One can see that the best SAM and ERGAS values are from the proposed S^{2}DBPN. Moreover, the UIQI value of S^{2}DBPN is also very close to the best UIQI value of 0.9591. Thus, the proposed method achieves better overall performance, which may benefit from the dual BP in spatial and spectral domains.

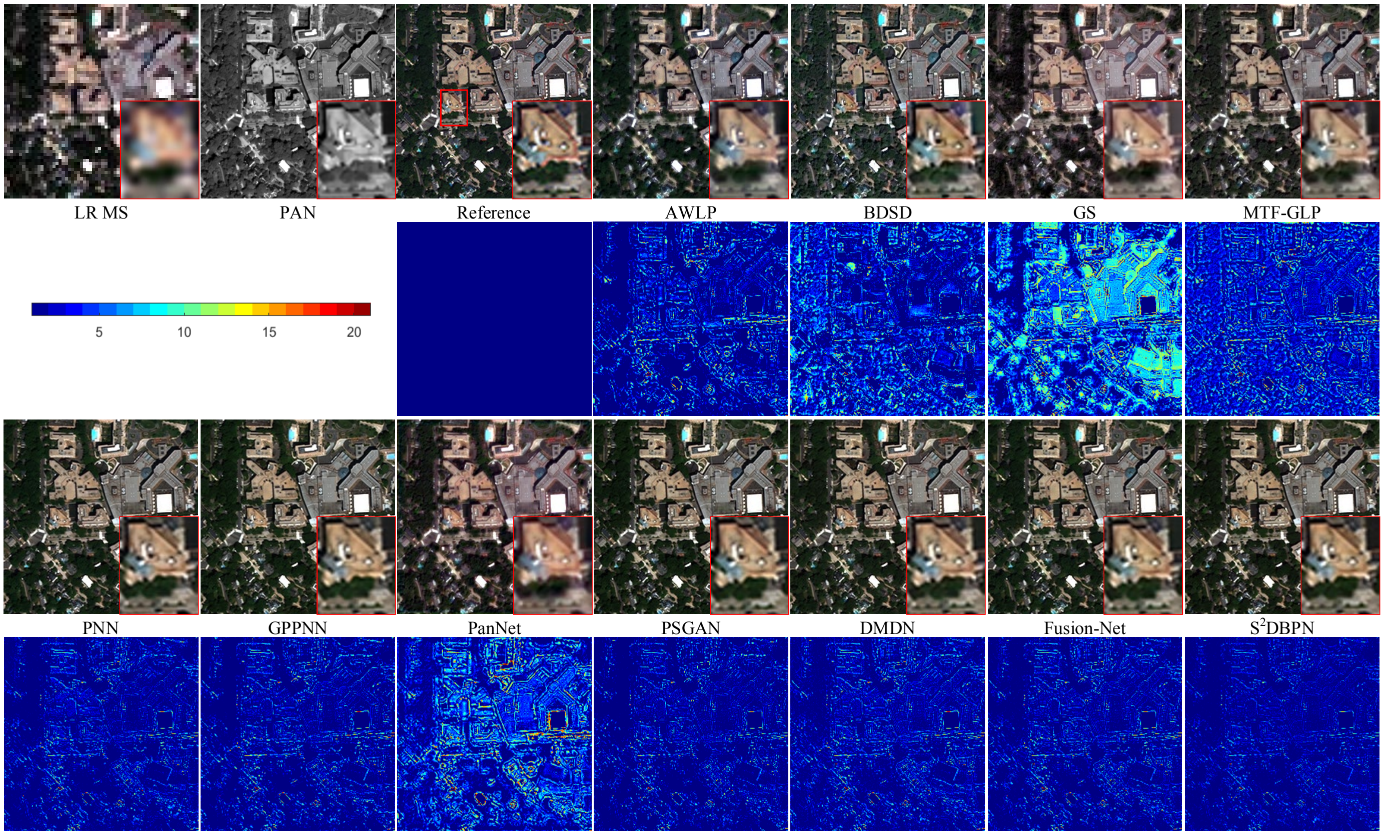


Fig. 8. Qualitative comparison of the fused results of all methods on the reduced-scale WorldView-2 dataset.

TABLE V. QUANTITATIVE COMPARISON ON THE REDUCED-SCALE DATASET FROM THE WORLDVIEW-2 SATELLITE.

Metric	AWLP	BDSD	GS	MTF-GLP	PNN	GPPNN	PanNet	PSGAN	DMDN	Fusion-Net	S^{2}DBPN
Q4	0.8871	0.7602	0.8761	0.8760	0.9487	0.9515	0.8709	0.9543	**0.9591**	0.9500	0.9590
RMSE	116.3824	162.1421	135.1632	118.6914	80.2298	70.4923	121.9907	74.0877	69.8947	69.9658	**69.8650**
SAM	6.8773	9.6111	7.3724	6.9648	5.3120	4.6880	6.9665	4.7541	4.7186	4.6742	**4.6518**
UIQI	0.8805	0.7531	0.7799	0.8673	0.9460	0.9557	0.8542	0.9539	0.9579	0.9572	**0.9584**
ERGAS	1.7428	2.2659	2.0643	1.7170	1.2289	1.1943	1.8022	1.1493	1.1161	1.1030	**1.1005**

C. Experiments on Full-Scale Datasets

This part illustrates the experimental results on full-scale QuickBird, GeoEye-1, and WorldView-2 datasets for comparison among all methods. Fig. 9 shows the fused images of all methods on the full-scale Quickbird dataset. An interesting area is chosen and enlarged for a more intuitive visual comparison. For traditional methods, the spatial details in fused images are enhanced well, but some spectral distortions appear in vegetation regions. The color of the tree in the AWLP and GS results is grey-green. There are slight differences between the fused images of PNN and GPPNN, especially in the building areas. For the PanNet result, the hue of the fused image is different from that of other fused images. In addition, the PSGAN result also contains some spectral distortions, which can be further noticed in the magnified region. Compared with the fused images of other methods, the S^{2}DBPN result preserves the spatial and spectral information in the fused image better.

Table VI presents the results of the fused images of different methods. 30 LR MS and PAN image pairs are tested and the average of metrics is computed. The proposed S^{2}DBPN has the best values in terms of D_λ and QNR. S^{2}DBPN obtains the second-best D_S, while the best value of D_S is from PNN.

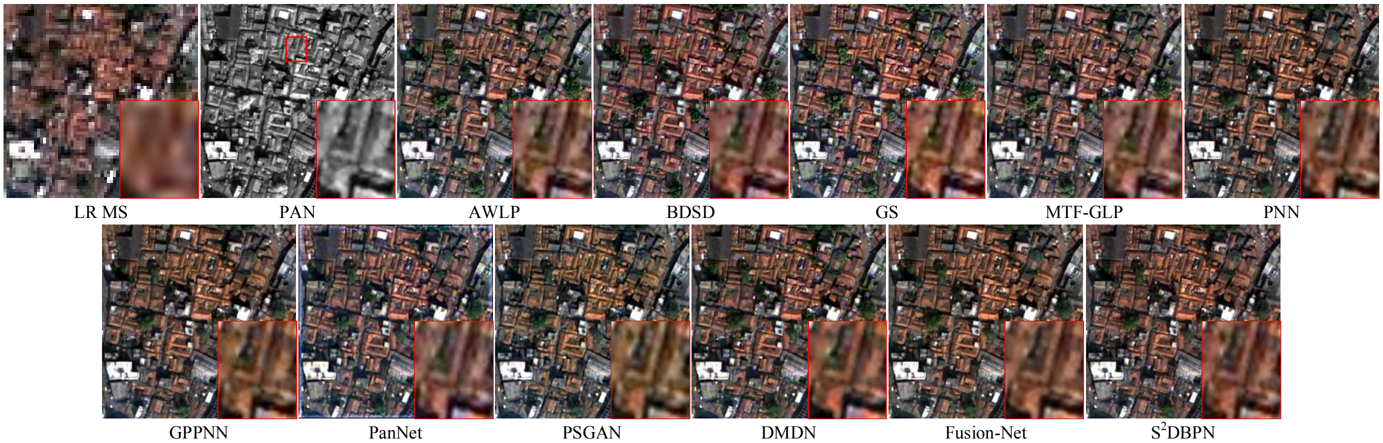


Fig. 9. Qualitative comparison of the fused results of all methods on the full-scale QuickBird dataset.

TABLE VI. QUANTITATIVE COMPARISON ON THE FULL-SCALE DATASET FROM THE QUICKBIRD SATELLITE.

Metric	AWLP	BDSD	GS	MTF-GLP	PNN	GPPNN	PanNet	PSGAN	DMDN	Fusion-Net	S^{2}DBPN
D_λ	0.0738	0.0294	0.0461	0.0630	0.0364	0.0364	0.0440	0.0352	0.0358	0.0335	**0.0261**
D_S	0.0555	0.0315	0.0385	0.0655	0.0232	0.0316	0.0272	0.0274	0.0268	**0.0222**	0.0271
QNR	0.8750	0.9400	0.9172	0.8760	0.9414	0.9333	0.9302	0.9383	0.9385	0.9451	**0.9474**

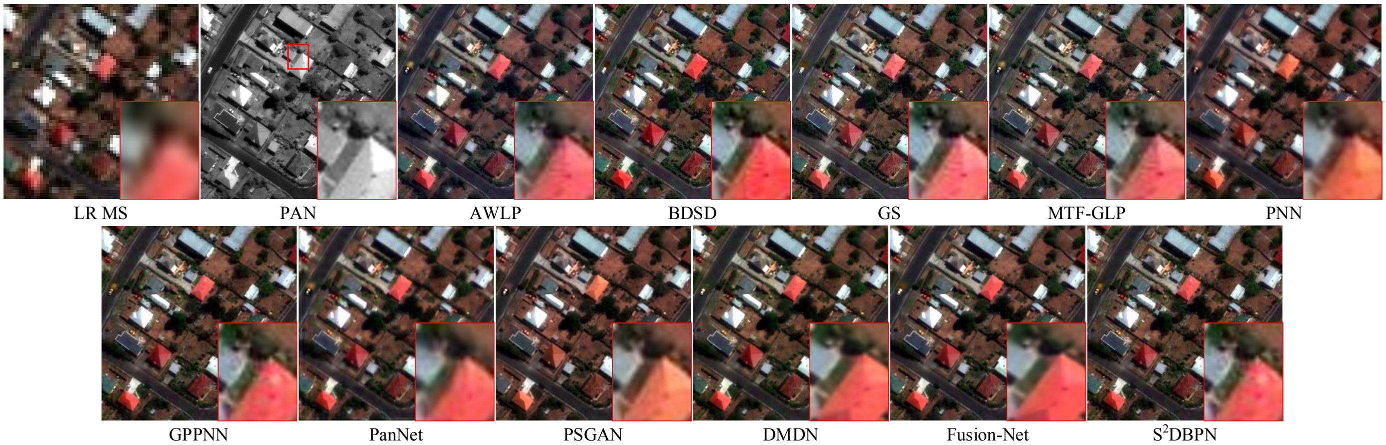


Fig. 10. Qualitative comparison of the fused results of all methods on the full-scale GeoEye-1 dataset.

TABLE VII. QUANTITATIVE COMPARISON ON THE FULL-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	AWLP	BDSD	GS	MTF-GLP	PNN	GPPNN	PanNet	PSGAN	DMDN	Fusion-Net	S^{2}DBPN
D_λ	0.1188	0.0457	0.0520	0.1158	0.0562	0.0592	0.0558	0.0693	0.0508	0.0658	**0.0472**
D_S	0.0471	0.0440	0.0415	0.0527	0.0545	**0.0340**	0.1036	0.0385	0.0384	0.0374	0.0396
QNR	0.8400	0.9126	0.9088	0.8381	0.8923	0.9089	0.8464	0.8950	0.9127	0.8994	**0.9151**

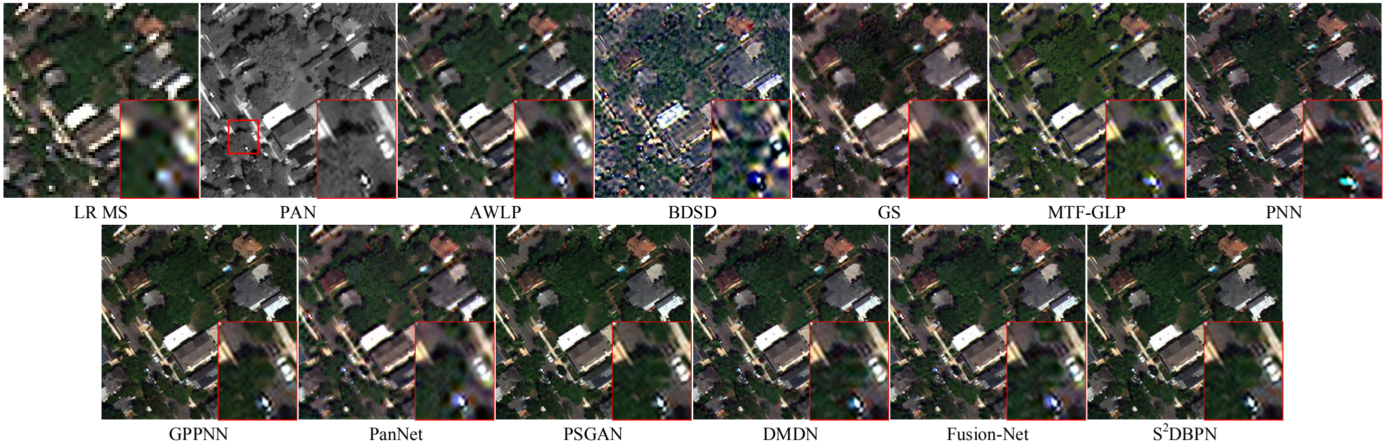
Fig. 10 demonstrates the fusion results of all methods on the full-scale GeoEye-1 dataset. The selected region is magnified and outlined by a red rectangle. The results on the GeoEye-1 dataset show more perceptible visual differences, especially in the magnified regions. For example, the color of the roof in the BDSD result is over-enhanced. But the color in the results of AWLP, GS, and MTF-GLP is in undersaturation, which may be caused by inappropriate transform coefficients in them. Spectral distortions also exist in the results of PNN and PSGAN. The results of GPPNN, PanNet, and S^{2}DBPN have a similar color. However, some spatial details in the PanNet result are lost.

Table VII lists the averaged values of each metric on 30 LR

MS and PAN image pairs. The best D_λ and QNR are provided by the proposed S^{2}DBPN. The D_S of S^{2}DBPN is close to the best one from GPPNN.

The fusion results of all methods on the full-scale WorldView-2 dataset are illustrated in Fig. 11. Compared to the results of other methods, significant spatial and spectral artifacts can be seen in the BDSD result. Moreover, the color information of the GS result is also not consistent with that of other fusion results. In the results of DNN-based methods, we can observe that the PanNet result is blurring, especially in the magnified area. In addition, some spectral artifacts appear in the magnified areas of DMDN and Fusion-Net results. The proposed S^{2}DBPN has better performance in terms of spectral and spatial preservation.

The quantitative results on the full-scale WorldView-2 dataset are listed in Table VIII. 30 LR MS and PAN image pairs are used for the test. The best D_S and QNR are produced by S^{2}DBPN. Thus, S^{2}DBPN balances the spatial and spectral information in fusion results better.


LR MS PAN AWLP BDSD GS MTF-GLP PNN

GPPNN PanNet PSGAN DMDN Fusion-Net S^{2}DBPN

Fig. 11. Qualitative comparison of the fused results of all methods on the full-scale WorldView-2 dataset.

TABLE VIII. QUANTITATIVE COMPARISON ON THE FULL-SCALE DATASET FROM THE WORLDVIEW-2 SATELLITE.

Metric	AWLP	BDSD	GS	MTF-GLP	PNN	GPPNN	PanNet	PSGAN	DMDN	Fusion-Net	S^{2}DBPN
D_λ	0.0655	0.0510	0.0425	0.0814	0.0536	0.0501	0.0536	0.0500	0.0468	**0.0392**	0.0470
D_S	0.0527	0.1299	0.1198	0.0830	0.0763	0.0522	0.0595	0.0506	0.0555	0.0570	**0.0448**
QNR	0.8854	0.8267	0.8434	0.8426	0.8741	0.9005	0.8906	0.9019	0.9004	0.9062	**0.9104**

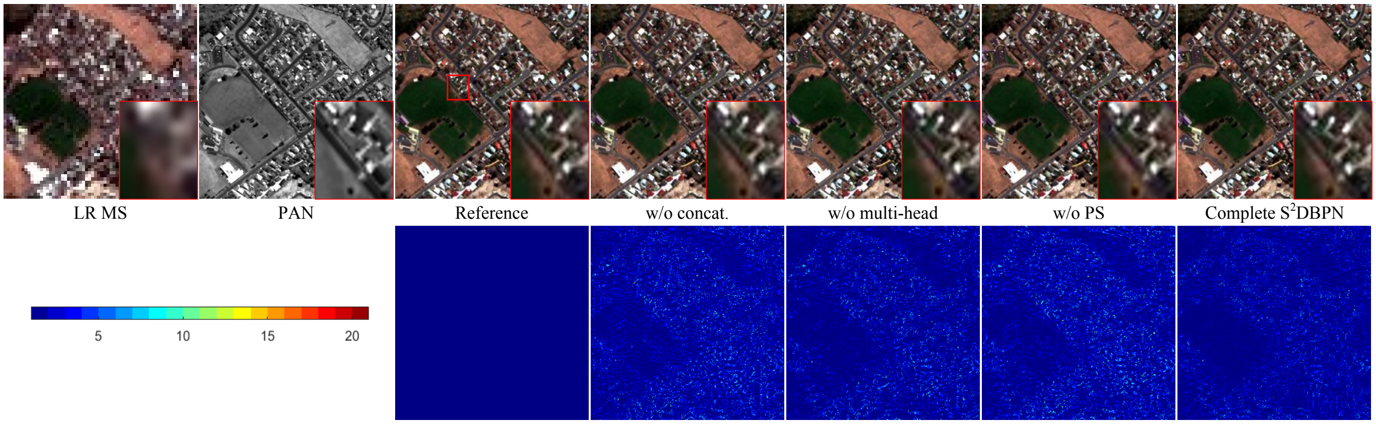

LR MS PAN Reference w/o concat. w/o multi-head w/o PS Complete S^{2}DBPN

Fig. 12. Ablation study of the proposed S^{2}DBPN on the reduced-scale GeoEye-1 dataset.

D. Ablation Study

To validate the effectiveness of S^{2}DBPN, different configurations are tested by removing specific modules. The ablation study is conducted on the reduced-scale GeoEye-1 dataset. Fig. 12 and Table IX demonstrate the qualitative and quantitative results of different configurations, respectively. First, we remove the concatenation of features from different stages in Fig. 1. The absolute difference maps show that the reconstruction errors become larger when the concatenation is removed. Second, the multi-head SpeUB is replaced by only one SpeUB. Without the multi-head mechanism, larger reconstruction errors can be found from the corresponding difference maps and the values of metrics in Table IX become inferior. In addition, the pixel shuffle (PS) is substituted by the bicubic interpolation operator. When the bicubic operation is adopted, some spatial artifacts are observed in the result of w/o PS because the up-sampling in the spatial up-projection module is adaptively learned by the PS operator. The results in Fig. 12 and Table IX imply that the introductions of these configurations boost the quality of the fused image of the proposed S^{2}DBPN.

TABLE IX. ABLATION STUDY ON THE REDUCED-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	w/o concat.	w/o multi-head	w/o PS	Complete S^2DBPN
Q4	0.8259	0.8263	0.8243	**0.8351**
RMSE	14.6316	14.8631	15.4657	**11.1062**
SAM	2.7925	2.8465	2.9763	**2.1052**
UIQI	0.9831	0.9826	0.9809	**0.9903**
ERGAS	0.9247	0.9394	0.9768	**0.7045**

E. *Network Architecture*

In this part, we explore the influence of the number of stages on the reduced-scale GeoEye-1 dataset. Fig. 13 shows the fusion results of the proposed S^2DBPN with various numbers of stages. Their absolute difference maps also give in the second row of Fig. 13. Table X provides the quantitative results on the reduced-scale GeoEye-1 dataset. In the first row of Fig. 13, it is difficult to distinguish the visual quality of these fused images. However, some differences can be observed from the second row of Fig. 13. It can be observed that the reconstruction errors are smaller when 3 stages are contained in S^2DBPN. In Table X, we also can find that the numerical values behave best for the S^2DBPN with 3 stages.

Besides, Table X also reports the number of parameters and training time of S^2DBPN with different numbers of stages. Because the weight-sharing strategy is adopted in S^2DBPN, the numbers of parameters are almost consistent. But it will spend more time to train the model by increasing the number of stages. Considering the fusion performance, the down- and up-projections in spatial and spectral domains are stacked three times.

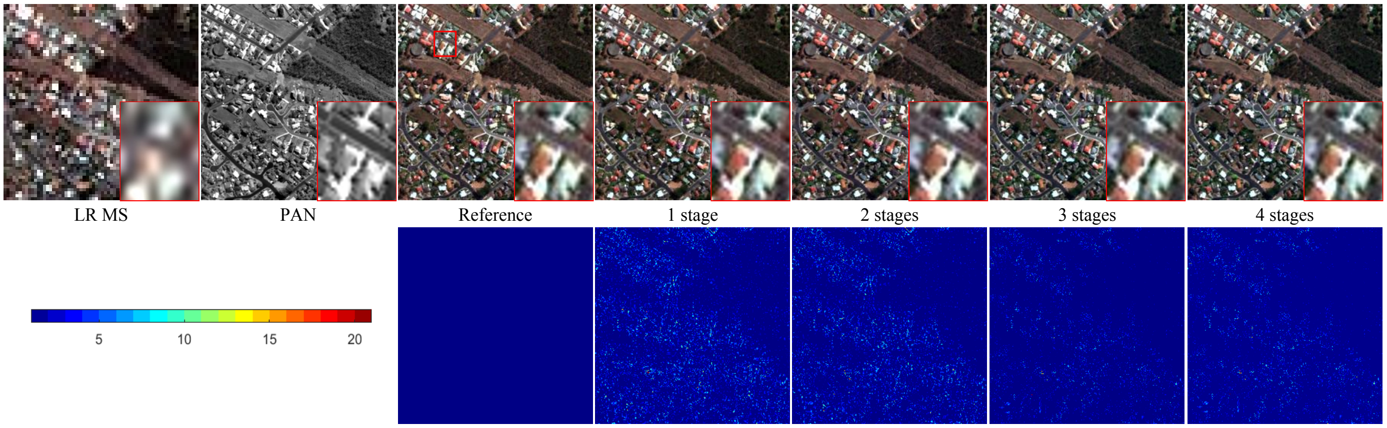


Fig. 13. Influences of the number of stages on the reduced-scale GeoEye-1 dataset.

TABLE X. INFLUENCES OF THE NUMBER OF STAGES ON THE REDUCED-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	1 stage	2 stages	3 stages	4 stages
Q4	0.8211	0.8252	**0.8351**	0.8265
RMSE	16.2461	14.9145	**11.1062**	14.3092
SAM	3.0667	2.8450	**2.1052**	2.7193
UIQI	0.9790	0.9824	**0.9903**	0.9839
ERGAS	1.0269	0.9426	**0.7045**	0.9047
#Para. (M)	**16.11**	16.13	16.15	16.16
#Time (h)	**32.1**	37.5	42.1	49.4

F. *Weight Sharing*

In the proposed S^2DBPN, the weight-sharing strategy is imposed on the three stages in Fig. 1 to reduce the model size. In this part, S^2DBPN with independent weights is trained and compared with the one that shares weights. The comparison is conducted on the reduced-scale GeoEye-1 dataset. Fig. 14 shows the fusion results and the corresponding absolute difference maps. The corresponding average results are given in Table XI. The difference maps in Fig. 14 suggest that the S^2DBPN with sharing weights can reconstruct the fused image better. The best values in Table XI are also produced by the S^2DBPN with sharing weights. The performance in Fig. 14 and Table VIII is similar to that in [43]. The reason for this may be that the S^2DBPN with independent weights cannot be trained sufficiently on the dataset used in this paper because its model size, 36.71M, is much greater than 16.15M of S^2DBPN with sharing weights as shown in Table XI. Owing to adopting the same structures, their computational complexities are comparable. Thus, their training time in Table XI is very close.

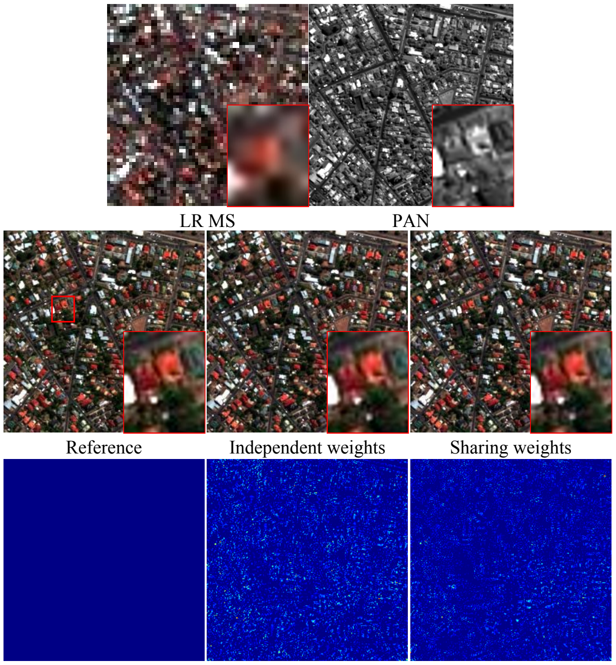


Fig. 14. Independent weights vs. sharing weights on the reduced-scale GeoEye-1 dataset.

TABLE XI. INDEPENDENT WEIGHTS VS. SHARING WEIGHTS ON THE REDUCED-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	Q4	RMSE	SAM	UIQI	ERGAS	#Para. (M)	#Time (h)
Independent weights	0.8272	14.3444	2.7163	0.9839	0.9075	36.71	**41.2**
Sharing weights	**0.8351**	**11.1062**	**2.1052**	**0.9903**	**0.7045**	**16.15**	42.1

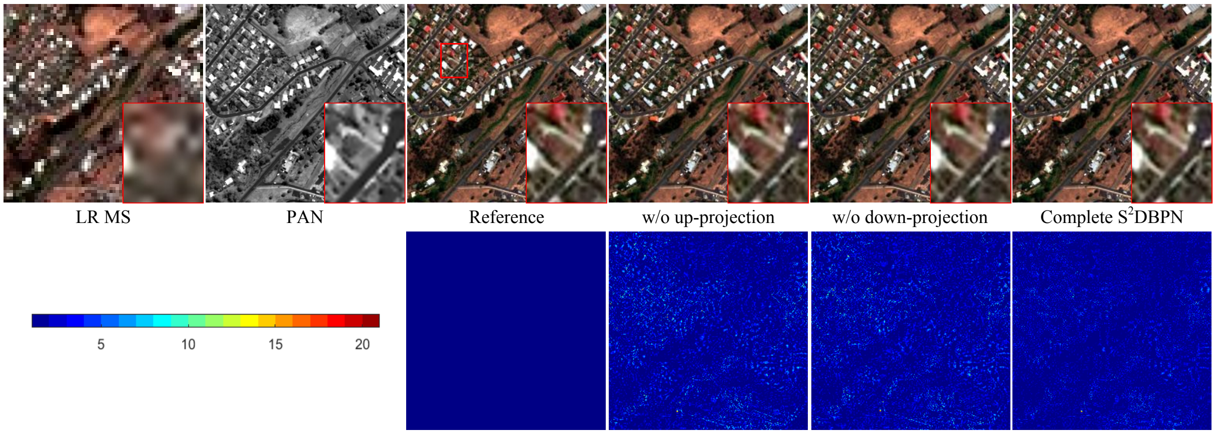


Fig. 15. Learnable projection VS. handcrafted projection on the reduced-scale GeoEye-1 dataset.

G. *Learnable Projection VS. Handcrafted Projection*

In this part, the learnable projection modules in Fig. 1 are replaced by the handcrafted projections to demonstrate the effectiveness of the modules composed of DNNs. The experiments are implemented on the reduced-scale GeoEye-1 dataset. The visual and numerical results are illustrated in Fig. 15 and Table XII, respectively. In Fig. 15, magnified regions and absolute difference maps are also shown for further analysis.

TABLE XII. LEARNABLE PROJECTION VS. HANDCRAFTED PROJECTION ON THE REDUCED-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	w/o learnable up-projection	w/o learnable down-projection	Complete S^{2}DBPN
Q4	0.8235	0.8279	**0.8351**
RMSE	15.6933	14.0762	**11.1062**
SAM	3.0327	2.6696	**2.1052**
UIQI	0.9803	0.9844	**0.9903**
ERGAS	0.9920	0.8895	**0.7045**

First, the spatial and spectral up-projection modules in S^{2}DBPN are replaced by the bicubic and linear interpolation operators, respectively. From Fig. 15, we can find spatial artifacts and larger reconstruction errors from the result of the S^{2}DBPN without learnable up-projection modules. Besides, we substitute the spatial and spectral down-projection modules in S^{2}DBPN with the handcrafted operators mentioned above. Similar performance is found in the result of the S^{2}DBPN without learnable down-projection modules. From the values in Table XII, one can see that the complete S^{2}DBPN produces the best values for all metrics. The results of S^{2}DBPN without learnable projection modules are inferior to those of the complete S^{2}DBPN. However, the S^{2}DBPN without learnable projection modules behaves better than the compared methods, as shown in Table IV. So, it validates the effectiveness of the BP-driven model.

H. *S^{2}DBPN VS. DBPN*

This part explores the effectiveness of the spatial-spectral dual BP formulation in S^{2}DBPN. Specifically, we directly used DBPN in [51] for comparison. In DBPN, only the spatial BP is contained and the spectral BP cannot be embedded into DBPN. So, DBPN is trained on the paired LR MS and HR MS images, without PAN images. From the fusion results in Fig. 16, we can see that the DBPN result is more blurring than the S^{2}DBPN result because the spatial information in PAN images is not considered in DBPN. The quantitative evaluations are given in Table XIII. Obvious performance degradation can be seen from the metric values of DBPN. The results in Table XIII also demonstrate that the spatial-spectral dual BP formulation can improve the quality of fused images better.

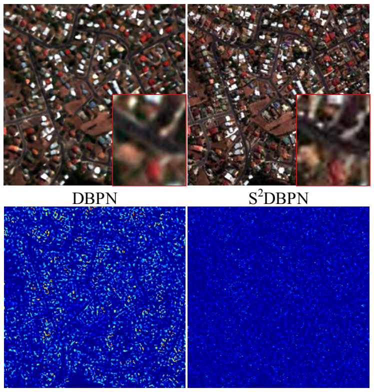


Fig. 16. S^{2}DBPN VS. DBPN on the reduced-scale GeoEye-1 dataset.

TABLE XIII. S^{2}DBPN VS. DBPN ON THE REDUCED-SCALE DATASET FROM THE GEOEYE-1 SATELLITE.

Metric	Q4	RMSE	SAM	UIQI	ERGAS
DBPN	0.7319	30.7993	4.0846	0.9242	1.9741
S^{2}DBPN	**0.8351**	**11.1062**	**2.1052**	**0.9903**	**0.7045**

V. CONCLUSION

In this paper, we have proposed S^{2}DBPN for the fusion of LR MS and PAN images, which is inspired by the BP model. Compared to the deep unfolding networks that are derived from sophisticated optimization algorithms, BP-driven DNNs are simple in principle, in which images are enhanced by up- and

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3266799

down-projections in the feature space. In the proposed S^{2}DBPN, the spatial BP network focus on the improvement of spatial information by successive spatial down- and up-projection modules. Spectral down- and up-projection modules are established according to the degradation relationship between PAN and HR MS images, by which the spectral BP is achieved. Then, the designed up- and down-projection modules are cascaded to enhance the spatial and spectral information in the fused image progressively. Finally, all features at different stages of spatial and spectral BP networks are combined to obtain the fused image. Experimental results on the datasets from QuickBird, GeoEye-1, and WorldView-2 satellites show the effectiveness of the proposed S^{2}DBPN compared with some state-of-the-art methods. Due to the cascaded down- and up-projection modules, the number of parameters of the proposed method increases significantly when more modules are introduced. So, the training time also becomes longer. For future work, we will design efficient and lightweight down- and up-projection modules and embed them into the dual BP framework to improve the reconstruction of fused images.

REFERENCES

[1] N. Liu, W. Li, Y. Wang, R. Tao, Q. Du, J. Chanussot, “A survey on hyperspectral image restoration: from the view of low-rank tensor approximation,” *Sci. China Inf. Sci.*, vol. 66, pp. 140302, Feb. 2023.

[2] N. Liu, L. Li, W. Li, R. Tao, J. Fowler, J. Chanussot, “Hyperspectral restoration and fusion with multispectral imagery via low-rank tensor-approximation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7817-7830, Sept. 2021.

[3] K. Zhang, F. Zhang, W. Wan, H. Yu, J. Sun, J. Del Ser, E. Elyan, A. Hussain, “Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead,” *Inf. Fusion*, vol. 93, pp. 227-242, 2023.

[4] G. Vivone, M. D. Mura, and A. Garzelli *et al.*, “A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods,” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53-81, Mar. 2021.

[5] X. Meng, Y. Xiong, F. Shao, H. Shen, W. Sun, and G. Yang, *et al.*, “A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation,” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 18-52, Mar. 2021.

[6] T. Tu, P. Huang, C. Hung, and C. Chang, “A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 309-312, Oct. 2004.

[7] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, T. Wittman, “An adaptive IHS pan-sharpening method,” *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746-750, Apr. 2010.

[8] B. Aiazzi, S. Baronti, and M. Selva, “Improving component substitution pansharpening through multivariate regression of MS+Pan data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230-3239, Oct. 2007

[9] K. Zhang, F. Zhang, Z. Feng, J. Sun, Q. Wu, “Fusion of panchromatic and multispectral images using multiscale convolution sparse decomposition,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 426-439, 2021.

[10] H. Shahdoosti, H. Ghassemian, “Combining the spectral PCA and spatial PCA fusion methods by an optimal filter,” *Inf. Fus.*, vol. 27, pp. 150-160, 2016.

[11] J. Duran, A. Buades, “Restoration of pansharpened images by conditional filtering in the PCA domain,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 442-446, Mar. 2019.

[12] A. Garzelli, F. Nencini, and L. Capobianco, “Optimal MMSE pan sharpening of very high resolution multispectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228-236, Jan. 2008.

[13] G. Vivone, Robust band-dependent spatial-detail approaches for panchromatic sharpening, *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421-6432, Sept. 2019.

[14] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, and J. Chanussot, “Contrast and error-based fusion schemes for multispectral image pansharpening,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 930-934, May 2014.

[15] X. Otazu, M. González-Audícana, O. Fors, and J. Nunez, “Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376-2385, Oct. 2005.

[16] H. Li, F. Liu, S. Yang, K. Zhang, X. Su, L. Jiao, “Refined pan-sharpening with NSCT and hierarchical sparse autoencoder,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5715-5725, Dec. 2016.

[17] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, “Remote sensing image fusion using the curvelet transform,” *Inform. Fus.*, vol. 8, no. 2, pp. 143-156, Sept. 2007.

[18] S. Zheng, W. Shi, J. Liu, and J. Tian, “Remote sensing image fusion using multiscale mapped LS-SVM,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1313-1322, May 2008.

[19] H. Yin, S. Li, “Pansharpening with multiscale normalized nonlocal means filter: A two-step approach,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5734-5745, Oct. 2015.

[20] Y. Xing, M. Wang, S. Yang, K. Zhang, “Pansharpening with multiscale geometric support tensor machine,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56 no. 5, pp. 2503-2517, May 2018.

[21] K. Zhang, M. Wang, S. Yang, and L. Jiao, “Convolution structure sparse coding for fusion of panchromatic and multispectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1117-1130, Feb. 2019.

[22] S. Li, H. Yin, and L. Fang, “Remote sensing image fusion via sparse representations over learned dictionaries,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4779-4789, Sept. 2013.

[23] F. Zhang, H. Zhang, K. Zhang, Y. Xing, J. Sun, Q. Wu, “Exploiting low-rank and sparse properties in strided convolution matrix for pansharpening,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2649-2661, 2021.

[24] S. Yang, K. Zhang, and M. Wang, “Learning low-rank decomposition for pan-sharpening with spatial-spectral offsets,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3647-3657, Aug. 2018

[25] R. Dian, S. Li, “Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization,” *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135-5146, Oct. 2019.

[26] P. Liu, L. Xiao, J. Zhang, B. Naz, “Spatial-hessian-feature-guided variational model for pan-sharpening,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2235-2253 Arp. 2016.

[27] P. Liu, L. Xiao, S. Tang, “A new geometry enforcing variational model for pan-sharpening,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5726-5739, Dec. 2016.

[28] L. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, “Machine learning in pansharpening: A benchmark, from shallow to deep networks,” *IEEE Trans. Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279-315, 2022.

[29] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, “Pansharpening by convolutional neural networks,” *Remote Sens.*, vol. 8, no. 7, pp. 594:1-594:22, July 2016.

[30] L. He, Y. Rao, J. Li, *et al.*, “Pansharpening via detail injection based convolutional neural networks,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188-1204, Apr. 2019.

[31] K. Zhang, A. Wang, F. Zhang, W. Diao, J. Sun, and L. Bruzzone, “Spatial and spectral extraction network with adaptive feature fusion for pansharpening,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5410814, 2022.

[32] S. Chen, H. Qi, K. Nan, “Pansharpening via super-resolution iterative residual network with a cross-scale learning strategy,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5407016, 2022.

[33] W. Huang, M. Ju, Q. Chen, B. Jin, W. Song, “Detail-injection-based multiscale asymmetric residual network for pansharpening,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 5512505, 2022.

[34] D. Lei, H. Chen, L. Zhang, and W. Li, “NLRNet: An efficient nonlocal attention ResNet for pansharpening,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5401113, 2022.

[35] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, Y. Zhang, “Residual encoder-decoder conditional generative adversarial network for pansharpening,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1573-1577, Sept. 2020.

[36] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A Generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 12, pp. 10227-10242, Dec. 2021.

[37] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, and L. Bruzzone, "ZeRGAN: Zero-reference GAN for fusion of multispectral and panchromatic images," *IEEE Trans. Neural Netw. Learn. Syst*., early access, Jan. 4, 2022.

[38] J. Li, W. Sun, M. Jiang, Q. Yuan, "Self-supervised pansharpening based on a cycle-consistent generative adversarial network," *IEEE Geosci. Remote Sens. Lett*., vol. 19, pp. 5511805, 2022.

[39] X. Meng, N. Wang, F. Shao, S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5409011, 2022.

[40] K. Zhang, Z. Li, F. Zhang, W. Wan, and J. Sun, "Pan-sharpening based on transformer with redundancy reduction," *IEEE Geosci. Remote Sens. Lett*., vol. 19, pp. 5513205, 2022.

[41] X. Sun, J. Li, Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5407423, 2022.

[42] F. Zhang, K. Zhang, J. Sun, "Multiscale spatial-spectral interaction transformer for pan-sharpening," *Remote Sens*., vol. 14, no. 7, pp. 1736, 2022.

[43] C. Mou, Q. Wang, J. Zhang, "Deep generalized unfolding networks for image restoration," in *Proc. IEEE CVPR*, Jun. 2022, pp. 17399-17410.

[44] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhan, "Deep gradient projection networks for pan-sharpening," in *Proc. IEEE CVPR*, Jun. 2021, pp. 1366-1375.

[45] G. Yang, M. Zhou, K. Yan, *et al*., "Memory-augmented deep conditional unfolding network for pan-sharpening," in *Proc. IEEE CVPR*, Jun. 2022, pp. 1788-1797.

[46] X. Cao, X. Fu, D. Hong, Z. Xu, and D. Meng, "PanCSC-Net: A model-driven deep unfolding method for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5404713, 2022.

[47] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 32, no. 3, pp. 1124-1135, Mar. 2021.

[48] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 29, no. 11, pp. 5345-5355, Nov. 2018.

[49] X. Tian, K. Li, Z. Wang, *et al*., "VP-Net: An interpretable deep network for variational pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5402716, 2022.

[50] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *J. Vis. Commun. Image Represen*., vol. 4, no. 4, pp. 324-335, 1993.

[51] M. Haris, G. Shakhnarovich, N. Ukita, "Deep back-projectinetworks for single image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 43, no. 12, pp. 4323-4337, Dec. 2021.

[52] M. Haris, G. Shakhnarovich, N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3897-3906.

[53] F. Zhu, Q. Zhao, "Efficient single image super-resolution via hybrid residual feature learning with compact back-projection network," in *Proc. IEEE ICCVW*, Oct. 2019, pp. 1-8.

[54] Z. Liu, L. Wang, C. Li, W. Siu, "Hierarchical back projection network for image super-resolution," in *Proc. IEEE CVPRW*, Jun. 2019, pp. 1-10.

[55] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE ICCV*, Oct. 2017, pp. 5449-5457.

[56] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 32, no. 5, pp. 2090-2104, 2021.

[57] L. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional networks for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 8, pp. 6995-7010, 2021.

[58] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett*., vol. 1, no. 4, pp. 313-317, Apr. 2004.

[59] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147-149.

[60] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett*., vol. 9, no. 3, pp. 81-84, Mar. 2002.

[61] L. Wald, "Quality of high resolution synthesized images: Is there a simple criterion?," in *Proc. 3rd Conf. Fusion Earth Data*, 2000, pp. 99-105.

[62] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens*., vol. 74, no. 2, pp. 193-200, Feb. 2008.