Deep Multiscale Detail Networks for Multiband Spectral Image Sharpening

Xueyang Fu, *Member, IEEE*, Wu Wang, Yue Huang, Xinghao Ding, *Member, IEEE*, and John Paisley, *Member, IEEE*

***Abstract*— We introduce a new deep detail network architecture with grouped multiscale dilated convolutions to sharpen images contain multiband spectral information. Specifically, our end-to-end network directly fuses low-resolution multispectral and panchromatic inputs to produce high-resolution multispectral results, which is the same goal of the pansharpening in remote sensing. The proposed network architecture is designed by utilizing our domain knowledge and considering the two aims of the pansharpening: spectral and spatial preservations. For spectral preservation, the up-sampled multispectral images are directly added to the output for lossless spectral information propagation. For spatial preservation, we train the proposed network in the high-frequency domain instead of the commonly used image domain. Different from conventional network structures, we remove pooling and batch normalization layers to preserve spatial information and improve generalization to new satellites, respectively. To effectively and efficiently obtain multiscale contextual features at a fine-grained level, we propose a grouped multiscale dilated network structure to enlarge the receptive fields for each network layer. This structure allows the network to capture multiscale representations without increasing the parameter burden and network complexity. These representations are finally utilized to reconstruct the residual images which contain spatial details of PAN. Our trained network is able to generalize different satellite images without the need for parameter tuning. Moreover, our model is a general framework, which can be directly used for other kinds of multiband spectral image sharpening, e.g., hyperspectral image sharpening. Experiments show that our model performs favorably against compared methods in terms of both qualitative and quantitative qualities.**

***Index Terms*— Deep learning, hyperspectral image (HSI) sharpening, image fusion, pansharpening, superresolution.**

I. Introduction

IMAGES with multiband spectra have been widely used for various applications, such as digital maps, agriculture, and mining. Compared with ordinary images with only one or a few bands, images with multiband spectra usually contain multiple bands of objects captured by satellites or sensors under different spectrums. However, due to the physical constraints, satellites or sensors often generate either a high-resolution (HR) single-band spectral panchromatic (PAN) image or several low-resolution multiband spectral (LRMS) images. To fully take advantage of all available information, the pansharpening, which is a widely concerned problem in multispectral image fusion, is usually adopted to simultaneously fuse the two components to produce an HR multispectral (HRMS) image. Another related application is the hyperspectral image (HSI) sharpening, which can be treated as a number of pansharpening subproblems. Fig. 1 shows an example to demonstrate the difference between the ordinary image superresolution and multiband spectral image sharpening. The conventional superresolution technology aims to map LR images to its visually pleasing HR version with the same band (channel) number. While the multiband spectral image sharpening is designed to fuse two inputs with different band numbers and generates a data cube, which contains the highest resolution in both spatial and spectral components. In this article, we design our model by mainly exploring the pansharpening problem.

Due to the powerful representation ability of deep convolutional neural networks, many researchers have utilized this technology for pansharpening. For example, pansharpening by deep neural networks [1] assumes that the relationship between HR/LR multispectral image patches is the same between the corresponding HR/LR PAN image patches, and uses this assumption to learn mapping relationships through neural networks. Pansharpening by convolutional neural networks (PNN) [2] modified the previous network architecture for superresolution [3], and augments the input by introducing nonlinear radiometric indices. However, the above-mentioned two deep learning-based methods simply treat the pansharpening as an image regression problem. While these methods achieve state-of-the-art results, they do not consider and explore the specific goals of the pansharpening, i.e., spectral and spatial preservations, but rather treat it as a black-box learning process. It is clear that, for pansharpening, spatial and spectral preservations are of crucial importance during fusion and should be focused on when learning a function mapping. This motivates us to propose a task-related approach based on deep neural networks.

Manuscript received June 28, 2019; revised February 14, 2020; accepted May 16, 2020. Date of publication June 2, 2020; date of current version May 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 81671766, Grant 81671674, Grant 61671309, Grant 61901433, and Grant U1605252, in part by the Fundamental Research Funds for the Central Universities under Grant 20720160075 and Grant 20720180059, in part by the CCF-Tencent Open Fund, and in part by the Natural Science Foundation of Fujian Province of China under Grant 2017J01126. *(Corresponding author: Xinghao Ding.)*

Xueyang Fu is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China.

Wu Wang, Yue Huang, and Xinghao Ding are with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: dxh@xmu.edu.cn).

John Paisley is with the Department of Electrical Engineering, Data Science Institute, Columbia University, New York, NY 10027 USA.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2996498

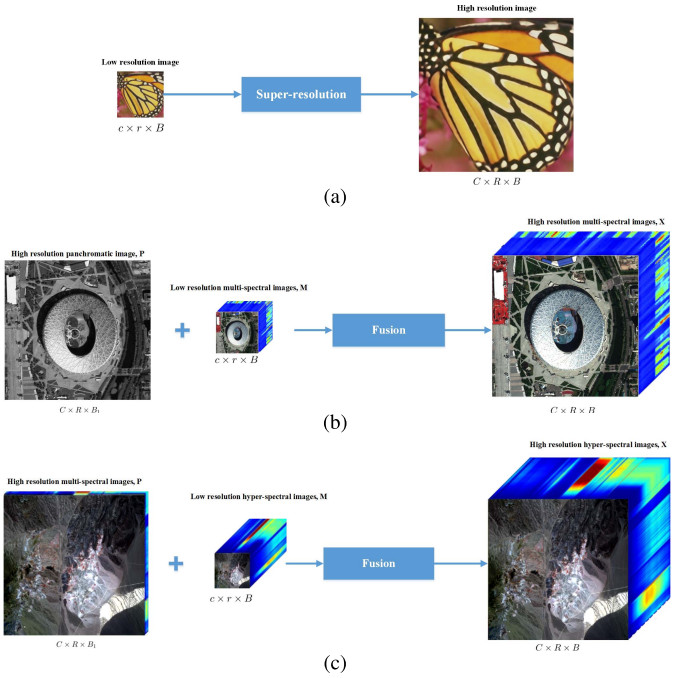


Fig. 1. Examples of conventional image superresolution and multiband spectral image sharpening. $C \times R$ and $c \times r$ represent the image size, B and B_1 represent the image band. (a) Image superresolution, $c < C$, $r < R$, and $B = 3$. (b) Pansharpening, $c < C$, $r < R$, and $1 = B_1 < B$. (c) HSI sharpening, $c < C$, $r < R$, and $1 < B_1 < B$.

In summary, our contributions are four-fold.

1) We incorporate our domain knowledge about multiband spectral image sharpening into the data-driven deep learning. Specifically, to preserve spectral information, we directly map the up-sampled multispectral images to the network output for a lossless propagation. To focus on spatial structures in the PAN image, we train the network in the high-frequency domain. We show that adding spectra mapping and using high-pass inputs can simplify the learning process.
2) We propose a new grouped multiscale dilated network structure to capture multiscale representations at a fine-grained level, which cannot be achieved by the widely used pooling operation. Our grouped multiscale structure is able to obtain larger receptive fields without increasing the calculation and storage burden.
3) We remove the batch normalization (BN) operation to improve generalization ability. We show that by doing so, our model can better handle different types of satellite images better than other deep learning-based methods.
4) Our network is a general model that can be directly used for multiband spectral image sharpening, e.g., pansharpening and HSI sharpening. Experiments on the two tasks show that our method is able to achieve state-of-the-art performance on both quantitative and qualitative qualities.

A preliminary version of this article called PanNet was presented earlier [4]. This article adds to the initial version in significant ways, and we summarize the changes in the following. First, compared with PanNet that only uses standard convolutions at a single scale, we add multiscale representation potential to extract rich contextual information. To achieve this goal, we propose a new grouped dilated block, which can obtain larger receptive fields than PanNet without significantly increasing the computational budget. Second, we argue that removing the BN, which is a basic module in PanNet, can further improve the generalization ability of the deep model while ensuring its sharpening performance. Third, we extend our model to HSI sharpening and provide more comprehensive evaluations and analysis.

II. Related Work

In recent decades, various pansharpening methods [5]–[8] have been proposed. Among these, the most popular methods are based on component substitution [9], including intensity hue-saturation technique [10], principal component analysis [11], and Brovey transform [12]. Though these methods are efficient and succeed in approximating the spatial resolution of the HRMS image, they tend to introduce spectral distortions.

More complex techniques have been proposed to address this problem, such as adaptive approaches (e.g., partial replacement adaptive component substitution (PRACS) [13]) and band-dependent approaches (e.g., band-dependent spatial-detail (BDSD) [14]). Multiresolution analysis methods have also been proposed [15]–[18]. In these approaches, the PAN and LRMS images are decomposed by using multiresolution tools, such as wavelets and Laplacian pyramids, and then fused. However, due to the differences in high-frequency regions, results generated by fusing these decomposed components usually contain aliasing and local dissimilarities.

Other handcrafted methods model the relationships between PAN, HRMS, and LRMS images by building regularized objective functions. These methods treat the fusion process as an image restoration optimization problem [19]–[24]. Since the spatial information is stored in PAN image, P+XS assumes that the spectral channels are contained in the topographic map of its PAN image. In other words, it considers that the PAN image is the linear combination of the HRMS image directly. The result of the P+XS method is remarkable, except it suffers from some blurry by penalizing large values [19]. In a recent study, a large part of methods introduces a high-pass filter to describe structural similarity while minimizing spectral distortion. Since the bands' range and pixel values of PAN image and HRMS image for the same objective are different, guided filter-based fusion [20], Bayesian nonparametric dictionary learning [21] and a regularized model-based optimization framework [22] consider the high-pass filtered components between the PAN image and the HRMS image is a linear relationship, and the error obey a Gaussian distribution. The recent method pansharpening with a hyper-Laplacian penalty (PHLP) [23] uses a hyper-Laplacian distribution to constrain the error, which allows large deviation of value in structural preservation to some degree. This penalty gives a significant improvement. In [24] and [25], the method achieves both satellite image registration and fusion (SIRF) in a unified framework, which not only utilizes a high-pass filter to achieve structural similarity but also incorporates the inherent correlation of different bands. PHLP and SIRF methods achieve

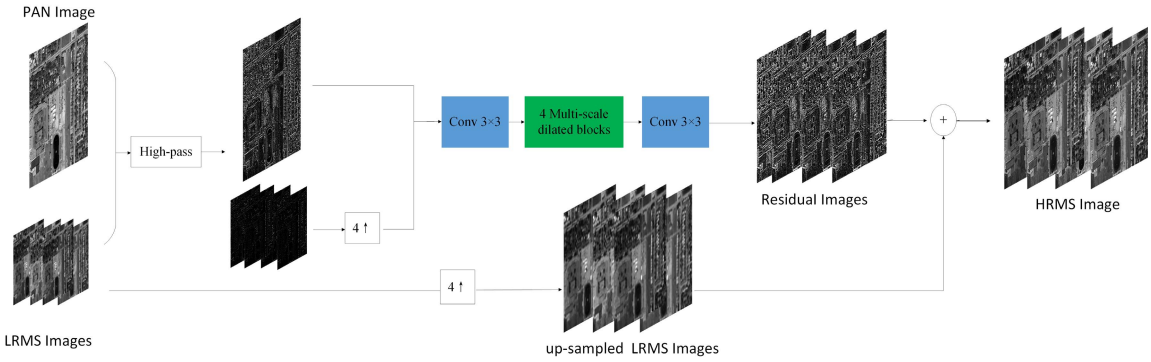


Fig. 2. Proposed deep network structure and multiscale convolutions for pansharpening. In each multiscale dilated block shown in Fig. 6, the input 64 feature maps are first divided into 4 groups for subsequent dilated convolutions. The final output features are obtained by fusing all multiscale features through the 1×1 convolution.

better results compared with previous methods. These methods can obtain excellent results, but also tend to depend on hand-designed assumptions, requiring parameter tuning for different satellites.

In the last few years, due to the powerful nonlinear modeling ability, data-driven deep convolution neural networks (CNNs) achieved remarkable success on both high-level vision tasks [26]–[30], [30] and low-level image processing problems [31], [32], such as image denoising [33]–[35], HSI sharpening [36], [37], superresolution [3], [38]–[40], compression artifacts reduction [41] biomedical image segmentation [42], and general image to image regression [43]. For pansharpening, Huang *et al.* [1] and Masi *et al.* [2] also take advantage of deep CNNs to tackle the fusion problem. However, as mentioned before, spectral and spatial preservations for pansharpening are not well considered in this two methods.

III. Proposed Multiscale Network

In general, the goal of pansharpening and HSI sharpening [44] is to utilize the HR component to sharpen the LR component that contains more bands and spectral information. Therefore, these two tasks can be unified into a single-observation model. We denote the desired images as $\mathbf{X}$ contains B bands with size of $C \times R$. The imaging models for the input images can be written as

$$\mathbf{P} = \mathbf{X}\mathbf{H}_\mathbf{p} + \mathbf{N}_\mathbf{p} \tag{1}$$

$$\mathbf{M} = \mathbf{H}_\mathbf{m}\mathbf{X} + \mathbf{N}_\mathbf{m} \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{CR \times B}$ is the desired output, $\mathbf{N}_\mathbf{p}$ and $\mathbf{N}_\mathbf{m}$ are the noises contained in the components $\mathbf{P}$ and $\mathbf{M}$, respectively. $\mathbf{H}_\mathbf{p} \in \mathbb{R}^{B \times B_1}$ is the response of the spectral sensor, and $\mathbf{H}_\mathbf{m} \in \mathbb{R}^{cr \times CR}$ is composed of a downsampling operator. Therefore, $\mathbf{P} \in \mathbb{R}^{CR \times B_1}$ is the spatial component that contains HR information with B_1 bands ($B_1 < B$), and $\mathbf{M} \in \mathbb{R}^{cr \times B}$ is the spectral component that contains B bands with size of $c \times r$ ($c < C$, $r < R$). To pansharpening, $\mathbf{P}$ is the HR PAN image and $\mathbf{M}$ is the LR multispectral images. To HSI sharpening, $\mathbf{P}$ is the HRMS images and $\mathbf{M}$ is the LR HSI s. As shown in Fig. 1, it is clear that each band of multispectral images in HSI sharpening plays the role of a PAN image in pansharpening. The HSI sharpening can be seen as a number of pansharpening subproblems. Therefore, in this article, we take the pansharpening problem as our main example since it can be directly extended to HSI sharpening.

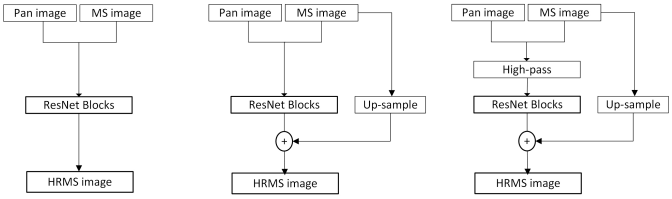


Fig. 3. Basic network structures we considered for pansharpening. From left to right: ResNet [26], ResNet + spectra mapping, and PanNet [4].

Fig. 2 shows the framework of our proposed deep learning with multiscale dilated convolution to pansharpening. As can be seen, this involves a skip connection between the LRMS image and output for enforcing spectral similarity and employing proposed multiscale dilated blocks to train network parameters in the high-pass domain for modeling spatial content. We first review common approaches to the pansharpening and then discuss our model in the context of the two goals of pansharpening, i.e., preserving spatial content of PAN image and spectral information of LRMS images.

A. Background

We denote the output HRMS images as $\mathbf{X}$ and the bth band image of $\mathbf{X}$ is $\mathbf{X}_b$. For the input images, $\mathbf{P}$ and $\mathbf{M}$ denote the PAN image and LRMS images, respectively. Most existing methods treat pansharpening as an optimization problem, which is often constructed from a Bayesian perspective by maximizing the posterior $\mathcal{P}(\mathbf{X}|\mathbf{P}, \mathbf{M})$. In general, maximizing the posterior can be transferred to minimize

$$\mathcal{L} = \lambda_1 J_1(\mathbf{X}, \mathbf{P}) + \lambda_2 J_2(\mathbf{X}, \mathbf{M}) + \lambda_3 J_3(\mathbf{X}) \tag{3}$$

where $J_1(\mathbf{X}, \mathbf{P})$, $J_2(\mathbf{X}, \mathbf{M})$, and $J_3(\mathbf{X})$ correspond to the structural consistency, spectral consistency, and the prior knowledge of $\mathbf{X}$, respectively. λ_1, λ_2 and λ_3 are regularization parameters. The desirable solution is the one that minimizes all the three terms. For example, the first variational method P+XS [19] uses linear assumptions to model structural consistency $J_1(\cdot)$

$$f_1(\mathbf{X}, \mathbf{P}) = \| \sum_{b=1}^{B} \omega_b \mathbf{X}_b - \mathbf{P} \|_F^2 \tag{4}$$

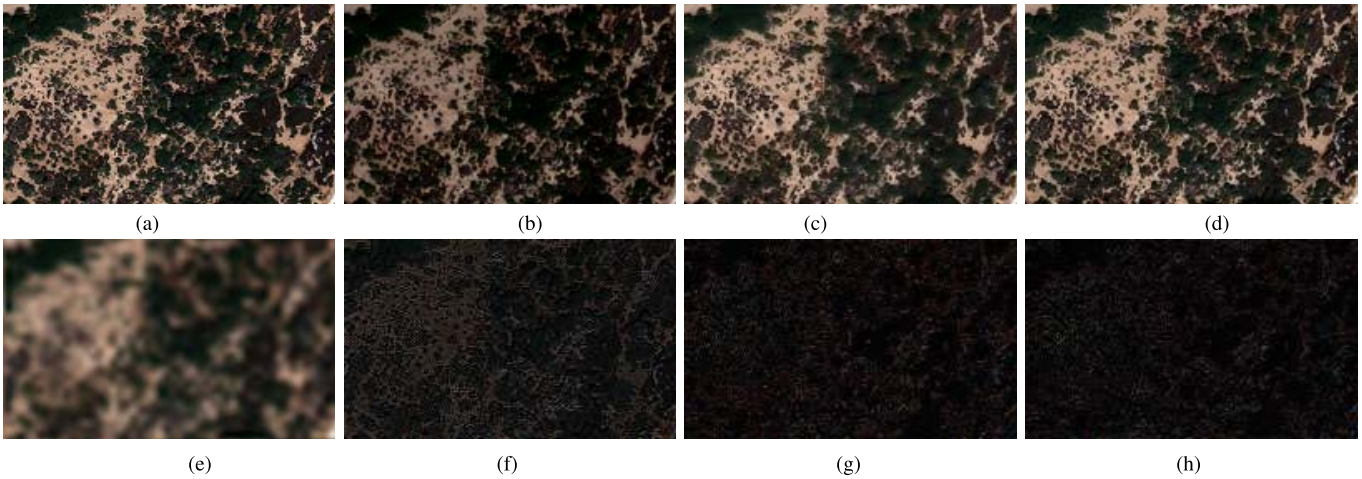


Fig. 4. Example result of the three model structures considered in Fig. 3. Better spectral modeling is observed in (c) and (d), while (d) has some spatial improvement over (c) in high-frequency details. (a) HRMS (ground truth). (b) ResNet [26]. (c) ResNet+spectra mapping. (d) PanNet [4]. (e) LRMS. (f) Residual of (b). (g) Residual of (c). (h) Residual of (d).

where ω is a B-dimensional probability weight vector. To focus on high-frequency parts, other methods utilize a spatial difference operator to preserve details. In [23], a hyper-Laplacian prior is further explored for structural penalty. For $J_2(\cdot)$, many methods define the spectral penalty as

$$f_2(\mathbf{X}, \mathbf{M}) = \sum_{b=1}^{B} \|\mathbf{k} \otimes \mathbf{X}_b - UP(\mathbf{M}_b)\|_F^2 \tag{5}$$

where $\otimes$ is the convolutional operation. $UP(\cdot)$ indicates the upsampling operation to make $\mathbf{M}_b$ to be the same size as $\mathbf{X}_b$, which is smoothed by the convolutional kernel $\mathbf{k}$ [19], [20], [22], [23]. While the most widely used penalization of $J_3(\cdot)$ is the TV regularization.

A straightforward strategy to utilize deep learning is to directly learn a mapping function between the two inputs $\{\mathbf{P}, \mathbf{M}\}$ and output $\mathbf{X}$. A simple and plain network architecture can be directly leveraged to minimizes

$$\mathcal{L} = \|f(\mathbf{P}, \mathbf{M}; \Theta) - \mathbf{X}\|_{\mathbf{F}}^2 \tag{6}$$

where $f(\cdot)$ represents a deep network and Θ denotes learnable network parameters. For example, this strategy is used by PNN [2], which inputs corresponding augmented training data into a deep CNN. Although this basic network achievers good results, it does not exploit characteristics of images used in pansharpening to define the network inputs or structures.

B. Basic Network Structure

We use the CNN with the ResNet block [26] as our network backbone. The deep CNN is able to capture relevant image characteristics and modeling complex nonlinear functions for regression tasks [26], [45]. Moreover, convolutional filters can also explore the high correlation across different multispectral image bands [24]. Therefore, to utilize the powerful nonlinear ability of deep structures, we adopt the popular ResNet as our basic network block. The overall architecture is expressed as

$$\begin{aligned} \mathbf{Y}^1 &= \sigma(\mathbf{W}^1 \otimes \text{concat}(\mathbf{P}_\text{G}, \text{UP}(\mathbf{M}_\text{G})) + \mathbf{v}^1) \\ \mathbf{Y}^{2l} &= \sigma(\mathbf{W}^{2l} \otimes \mathbf{Y}^{2l-1} + \mathbf{v}^{2l}) \\ \mathbf{Y}^{2l+1} &= \sigma(\mathbf{W}^{2l+1} \otimes \mathbf{Y}^{2l} + \mathbf{v}^{2l+1}) + \mathbf{Y}^{2l-1} \end{aligned} \tag{7}$$

where G denotes the high-pass information, $\sigma(\cdot)$ is the nonlinear operation, $\mathbf{W}$ and $\mathbf{v}$ denote the weights and biases, respectively. $l = 1, \ldots, (L-2)/2$, $\mathbf{Y}^l$ represents the lth layer output, $\uparrow$ represents the upsampling operation, $\mathbf{P}_G$, and $\text{UP}(\mathbf{M}_G)$ are directly concatenated and represented by the function $concat(\cdot)$. The final prediction $\hat{\mathbf{X}}$ is

$$\hat{\mathbf{X}} \approx (\mathbf{W}^L \otimes \mathbf{Y}^{L-1} + \mathbf{v}^L) + \text{UP}(\mathbf{M}) \tag{8}$$

where $\text{UP}(\mathbf{M})$ represents the upsampled LRMS image.

We test three potential basic network structures of which the structures are shown in Fig. 3. The third network, i.e., PanNet [4], achieves the best performance. The first structure corresponds to directly applying plain ResNet to the pansharpening problem. Based on this network structure, we propose a novel model for pansharpening to preserve both spectral and spatial information, which we discuss in the following.

1) Spectral Preservation: For spectral preservation, we upsample $\mathbf{M}$ and use a skip connection to the deep network

$$\mathcal{L} = \|f(\mathbf{P}, \mathbf{M}; \Theta) + \text{UP}(\mathbf{M}) - \mathbf{X}\|_{\mathbf{F}}^2. \tag{9}$$

This term is inspired by the variational methods represented by (5), and it enforces that $\mathbf{X}$ shares the spectral content of $\mathbf{M}$. However, different from variational methods that utilize a smoothing kernel to convolve $\mathbf{X}$, we allow the deep network to automatically correct the HR differences. The second network in Fig. 3 corresponds to (9), and we call this network as "ResNet + spectra mapping." For PanNet [4], we include this spectra mapping and a modification of ResNet discussed next.

2) Structural Preservation: As discussed in Section III-A, to enforce structural consistency, most variational methods utilize the high-pass information contained in the PAN image. These methods are able to generate clearer details than the

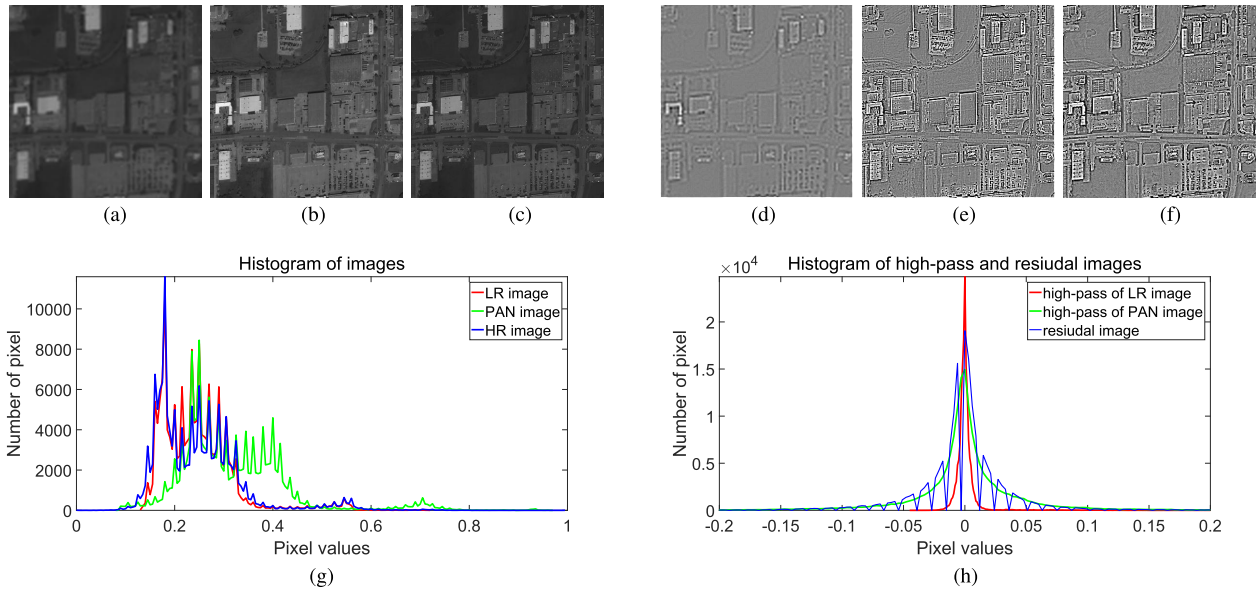


Fig. 5. Sparsity of high-pass components. Note that we only show one band of LR and HR images for the sake of demonstration. Other bands show similar phenomenons, i.e., high-pass components have the characteristics of sparsity during our experiments. (a) LR image. (b) PAN image. (c) HR image. (d) High pass of (a). (e) High pass of (b). (f) Residual (c)–(a). (g) Histogram of (a)–(c). (h) Histogram of (d)–(f).

P+XS approach, in which the image is directly used in (4). Based on this motivation, we use the high-pass content of the PAN image and of the upsampled LRMS images as the inputs of network. The modified model is

$$\mathcal{L} = \| f(\mathbf{P}_G, \mathrm{UP}(\mathbf{M}_G); \Theta) + \mathrm{UP}(\mathbf{M}) - \mathbf{X} \|_{\mathbf{F}}^{\mathbf{2}}. \tag{10}$$

To obtain the high-pass information, we subtract from the original images the low-pass content found by using the average filtering method. After obtaining the high-pass content, we upsample to the size of the PAN for LRMS images. Note that since $\mathrm{UP}(\mathbf{M})$ is the low-pass part, the term $(\mathrm{UP}(\mathbf{M}) - \mathbf{X})$ contains the high-pass component of $\mathbf{X}$. This frees the deep network to learn a mapping function that fuses the high-pass spatial information contained in PAN into $\mathbf{X}$. To make the network focus on processing high-pass information, we feed the high pass of $\mathrm{UP}(\mathbf{M})$, i.e., $\mathrm{UP}(\mathbf{M}_G)$, into the network.

In Fig. 4, we compare the networks shown in Fig. 3. Fig. 4(b) corresponds to the objective (6), Fig. 4(c) only considers the spectra mapping in (9), and Fig. 4(d) corresponds to (10). It is clear that spectra mapping focuses on spectral consistency and training on the high-frequency domain can well preserve edges and details.

C. Domain-Specific Knowledge for Learning Process

In this section, we qualitatively analyze how our domain-specific knowledge-based network design simplifies the learning process. As shown in Fig. 5, after adding spectra mapping and using high-pass inputs, the mapping process is actually between three sparse components, i.e., Fig. 5(d)–(f). In other words, most pixels equal to or close to 0, as shown in the histogram in Fig. 5(h). This indicates that the numbers of unknowns are significantly decreased, which makes the learning process easier to tackle.

Utilizing the sparsity is also widely used in existing pansharpening methods [20], [23]–[25]. Thus, based on our domain-specific knowledge, we introduce spectra mapping and high-pass inputs to train network parameters. As shown in Fig. 16, compared with ResNet and ResNet + spectral mapping, PanNet [4] has lower training and testing errors with the same network depth and training data, which demonstrates our design can simplify the learning process.

D. Grouped Multiscale Network Structure

Different from high-level vision problems, the pansharpening is an image fusion problem, which requires accurate dense pixel prediction. Thus, introducing the pooling operation, which is widely used to obtain abstract features, leads to spatial information loss that cannot be recovered. However, removing the pooling operation slows the increasing rate of the receptive field. On the other hand, many multiscale networks, e.g., U-Nets [42] and RBDN [43], in which features at lower scales, reuse the computation of features a higher scales using cleverly designed skip connections. However, these methods extract multiscale features in a layerwise fashion. The multiscale representation ability at a more granular scale is limited. Since we use the high-pass components as inputs, only fine details and edges are fed into the network. Therefore, to achieve the balance between spatial high-pass information preservation and receptive field enlargement, we propose a grouped multiscale dilated block to extract multiscale representations at a fine-grained level.

By weighting pixels with a step size of a dilated factor, the dilated convolution [46] can effectively increase the receptive field without losing spatial information and increasing parameter burdens. With different dilated factors, one fixed convolutional kernel is able to achieve various receptive fields, which inspire us to design a multiscale dilated block to fully utilize spatial information at different scales. However, directly increasing dilated factors require more computational budget. Therefore, we divide the feature maps into small groups

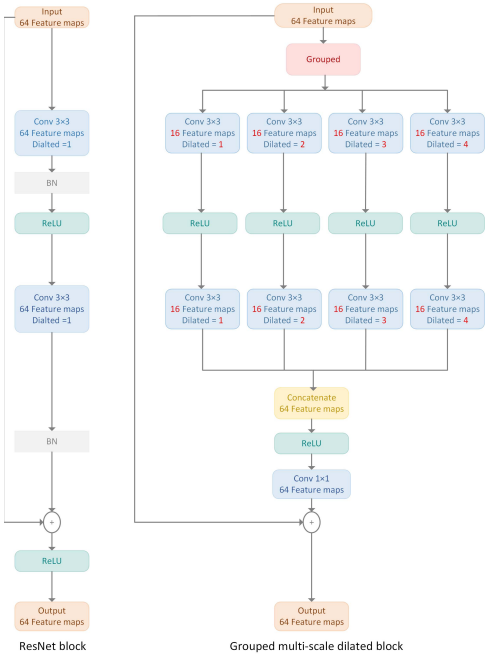


Fig. 6. Insights of ResNet block and our multiscale dilated block. d indicates the current dth layer. BN and ReLU indicate BN and rectifier linear unit, respectively.

and perform different dilated convolutions for each group separately.

Fig. 6 shows the insight of ResNet block and our grouped multiscale dilated block. The convolution operation in the ResNet block can be seen as a dilated convolution with a dilated factor equals to 1. Our grouped multiscale dilated block is consisted of two multiscale dilated operations and one 1×1 convolution. Each block contains four parallel dilated convolutions with different dilated factors. These four paths are concatenated and fused by one 1×1 convolution layer, which introduce negligible computational complexity, to generate the output feature maps. Note that the four paths are individually processed and not fully connected, which saves both computational and storage burdens. The multiscale representation ability of our grouped block is orthogonal to existing methods [42], [43] that utilizes features with different resolutions at different network layers. The multiscale of our model refers to the multiple receptive fields at a single network layer. Our whole network contains four grouped multiscale dilated blocks and two 3×3 convolution layers, as shown in Fig. 2. The first 3×3 convolution layer is used to extract basic image features while the last one is used for reconstructing the residual images.

E. Removing Batch Normalization

As one of the most effective way to alleviate the internal covariate shift, BN [47] is widely adopted before the nonlinearity in each layer in existing deep learning based methods. The operation of BN contains two parts: first, the feature maps x within a mini-batch are normalized by

$$\hat{x} = \frac{x - \mu_x}{\sigma_x} \tag{11}$$

where μ_x and σ_x are the mean and standard deviation of x within the mini batch. Then, the $\hat{x}$ is scaled and shifted by

$$y = \gamma \hat{x} + \beta \tag{12}$$

where y is the output feature maps, γ and β are learnable parameters and used to increase the representation ability.

The BN has several merits, such as fast training and low sensitivity to initialization. However, during our experiments, we found that BN does not always perform well when testing other satellite data. This is because BN assumes that the distributions of training and testing data are the same. In the test phase, the mean μ_x and the standard deviation σ_x used in (11) are calculated and saved based on the training data. However, for the remote sensing community, different satellites have their own data types. The μ_x and σ_x obtained on one satellite data are not always consistent with other satellites. Therefore, when testing new satellite data, the parameters learned from different satellites may cause distribution fluctuations and affect subsequent calculations. Note that, as described in PanNet, since the main energy in the image, i.e., the low-pass components, has been removed, training the network using high-pass details can reduce the distribution difference to some extent. However, continuous use of (11) and (12) will lead to the accumulation of fluctuations, which will increase the distribution difference again. Therefore, we argue that by combining with our high-pass details training, removing BN can further improve the generalization ability of deep networks for different satellites. This has practical value in the case of new satellites and sensors that cannot provide sufficient training data.

In addition, introducing spectra mapping and high-pass details can effectively simplify the learning process, as described in Section III-C. This implies that we do not need BN to accelerate training since the problem already becomes easy to handle. Moreover, removing BN can sufficiently reduce memory usage since the BN layers consume the same amount of memory as the preceding convolutional layers. Based on the above-mentioned observation and analysis, we remove BN layers from our network to improve generalization to new satellites and reduce parameter numbers and computing resources.

IV. Experiments

We conduct the experiments by using the Worldview3 satellite images. Since the HRMS images are not available in the data set, we follow Wald's protocol [48] for all experiments. The Wald protocol downsamples both the LRMS and PAN images so that the original LRMS images can be used as the ground truth images. Before downsampling, a low-pass filter is applied to reduce aliasing. To match the sensor properties, we follow the method [18] and use an approximation of the sensor modulation transfer function. Specifically, we use a 7×7 Gaussian kernel with a standard deviation of 0.1 to convolve all original images followed by downsampling with a factor of 4. We compare with several nondeep learning-based pansharpening methods: "à trous" wavelet transform (ATWT)-M3 [17], additive wavelet luminance proportional (AWLP) [16], BDSD [14], PRACS [13],

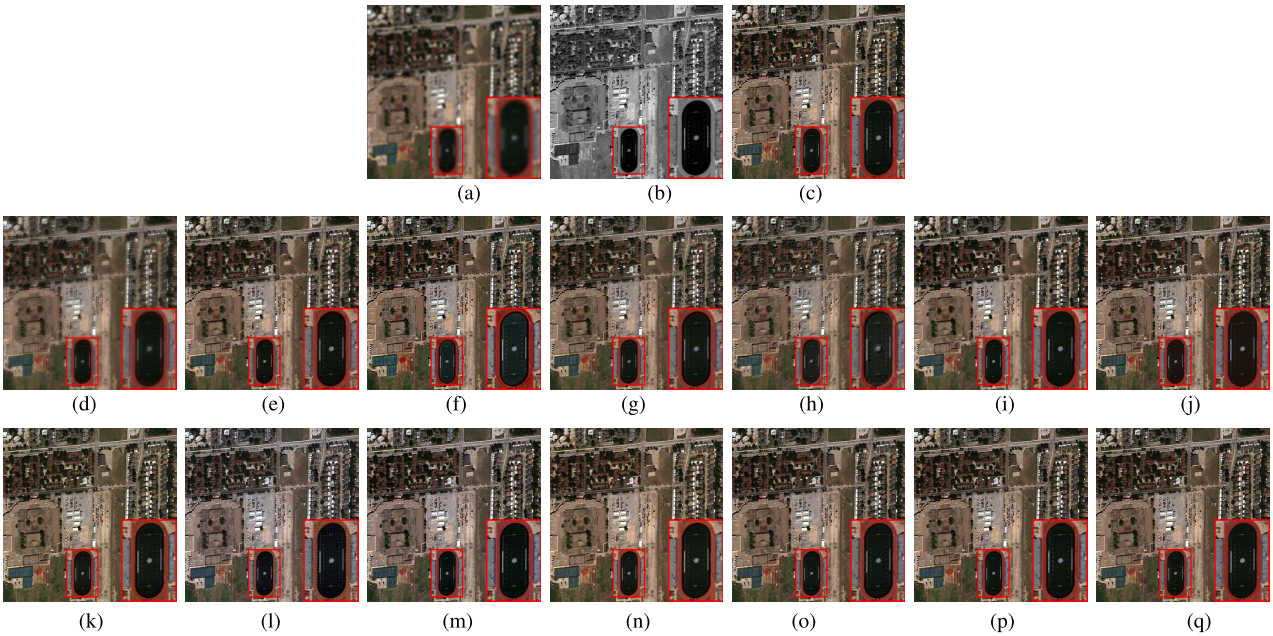


Fig. 7. Visual comparisons on the simulated data. As shown in the red rectangles, our method can simultaneously achieve both spatial and spectral preservation, e.g., there are fewer blurry artifacts near the track's edge, and it is closer to the color of the ground truth. Please zoomed-in view for a better visualization. (a) LRMS. (b) PAN. (c) Ground truth. (d) ATWT-M3. (e) AWLP. (f) BDSD. (g) PRACS. (h) Indusion. (i) PHLP. (j) SIRF. (k) PNN. (l) U-Nets. (m) RBDN. (n) ResNet. (o) ResNet + spec. (p) PanNet. (q) Our.

Indusion [15], PHLP [23], and SIRF [24], [25]. We compare with three state-of-the-art deep learning-based methods: one relative shallow network PNN [2] and two multiscale networks U-Nets [42] and RBDN [43]. We also compare with the other three network structures of Fig. 3, i.e., ResNet [26], ResNet + spectra mapping and PanNet [4].

To train our network, we totally extract $18K$ PAN/LRMS/HRMS patch pairs, and each patch size is set as 64×64. During the training process, 90% of the pairs are used to learn the network, and the rest is used for testing. We use the Caffe [49] to train our models and select ReLU [50] as the nonlinearity $\sigma(\cdot)$. The number of filters is set as 16 for all layers. We use the SGD algorithm, in which the weight decay and momentum are set as 10^{-7} and 0.9, to minimize the objective function (10). The learning rate is initialized as 0.001 and divided by 10 after 10^5 and 2×10^5 iterations. The training is finished at 2.5×10^5 iterations. The mini-batch size is 16, and the radius of the low-pass filter is 5.

A. Simulated Experimental Results

We first test our model on 225 images, which contain 8 spectral bands, from the Worldview3 satellite by using the experimental framework described earlier. Note that we only display the three color bands for visualization, while all spectral bands are used to perform quantitative evaluations. Five widely used quantitative metrics are utilized to evaluate performance, i.e., relative dimensionless global error in synthesis (ERGAS) [51], spectral angle mapper (SAM) [52], universal image quality index [53] averaged over the bands (QAVE) and X-band extension of Q8 (for 8 bands) [54], and the spatial correlation coefficient (SCC) [55].

The mean and standard deviation of quantitative scores are shown in Table I. It can be seen that if not considering the networks of Fig. 3 and our multiscale network, the PNN method performs the best. While PanNet [4] significantly improves the results over PNN [2]. This is due to the additional design of spectra-mapping and high-frequency inputs. Furthermore, our multiscale network achieves the best performance over all other methods, which indicates that using a multiscale fashion can further improve the reconstruction accuracy. This is because more contextual information is utilized for the subsequent reconstruction.

In Fig. 7, we show an example at the reduced scale. As shown in the red rectangles, other compared methods have obvious blurring and artifacts in their results and some spectral distortions (here showing as color distortion). In Fig. 8, we show the residuals of these images to highlight the differences. As can be seen, the color of the residual image of our multiscale network tends to gray, which means good spectra preservation. Meanwhile, our residual image also shows less detail and texture than other methods, which means the best spatial preservation is achieved by our model.

In Table II, we show the comparison on trainable parameter numbers of deep learning-based methods. As can be seen, our network has a comparable parameter number with PanNet [4] and much less than other methods, while achieves the best pansharpening performance, as shown in Table I.

B. Evaluation at the Original Scale

We also evaluate the results of the different methods at the original resolution of the WorldView3 satellite on 200 test images. One example of the results is shown in Fig. 9. Since we are lack of the ground truth HRMS images, the residuals

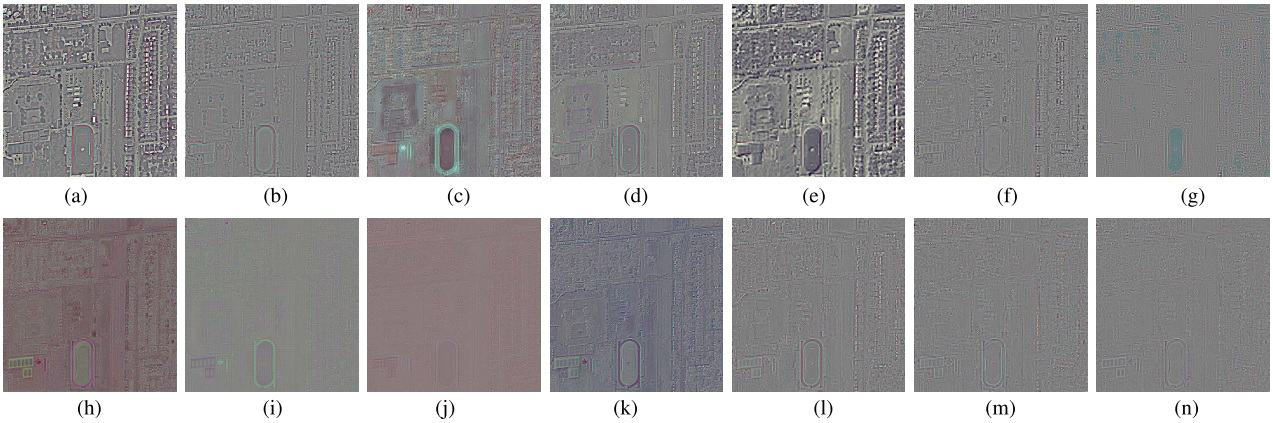


Fig. 8. Residuals between the HRMS image reconstructions and the ground truth from Fig. 7. (a) ATWT-M3. (b) AWLP. (c) BDSD. (d) PRACS. (e) Indusion. (f) PHLP. (g) SIRF. (h) PNN. (i) U-Nets. (j) RBDN. (k) ResNet. (l) ResNet + spec. (m) PanNet. (n) Our.

TABLE I

QUALITY METRICS OF DIFFERENT METHODS ON 225 SATELLITE IMAGES WORLDVIEW3. THE BEST AND THE SECOND BEST RESULTS ARE BOLDFACED AND UNDERLINED, RESPECTIVELY

Methods	Q8	QAVE	SAM	ERGAS	SCC
ATWT-M3 [17]	0.750 ± 0.025	0.741 ± 0.028	7.090 ± 1.808	4.693 ± 0.847	0.811 ± 0.036
AWLP [16]	0.710 ± 0.047	0.710 ± 0.048	6.946 ± 1.774	4.858 ± 0.920	0.798 ± 0.033
BDSD [14]	0.871 ± 0.030	0.867 ± 0.033	6.675 ± 1.909	3.631 ± 0.848	0.856 ± 0.048
PRACS [13]	0.836 ± 0.027	0.822 ± 0.030	6.385 ± 1.628	3.834 ± 0.796	0.835 ± 0.050
Indusion [15]	0.799 ± 0.025	0.799 ± 0.027	7.087 ± 1.544	4.340 ± 0.787	0.825 ± 0.032
PHLP [23]	0.859 ± 0.033	0.835 ± 0.042	5.748 ± 1.258	3.747 ± 0.587	0.845 ± 0.025
SIRF [24], [25]	0.863 ± 0.030	0.859 ± 0.030	5.277 ± 1.416	3.564 ± 0.642	0.866 ± 0.029
PNN [2]	0.882 ± 0.034	0.891 ± 0.025	4.752 ± 0.870	3.277 ± 0.472	0.915 ± 0.015
U-Nets [42]	0.837 ± 0.032	0.853 ± 0.021	6.107 ± 0.784	3.923 ± 0.415	0.887 ± 0.014
RNBN [43]	0.929 ± 0.029	0.931 ± 0.034	4.151 ± 0.691	2.722 ± 0.382	0.950 ± 0.010
ResNet [26]	0.847 ± 0.034	0.886 ± 0.028	4.940 ± 0.941	3.838 ± 0.495	0.917 ± 0.018
ResNet+spectra-map	0.905 ± 0.025	0.905 ± 0.026	4.730 ± 0.959	2.933 ± 0.498	0.918 ± 0.016
PanNet [4]	0.925 ± 0.023	0.928 ± 0.024	4.128 ± 0.787	2.469 ± 0.412	0.943 ± 0.012
Our	**0.934 ± 0.029**	**0.935 ± 0.032**	**3.681 ± 0.684**	**2.292 ± 0.373**	**0.957 ± 0.009**
ideal value	1	1	0	0	1

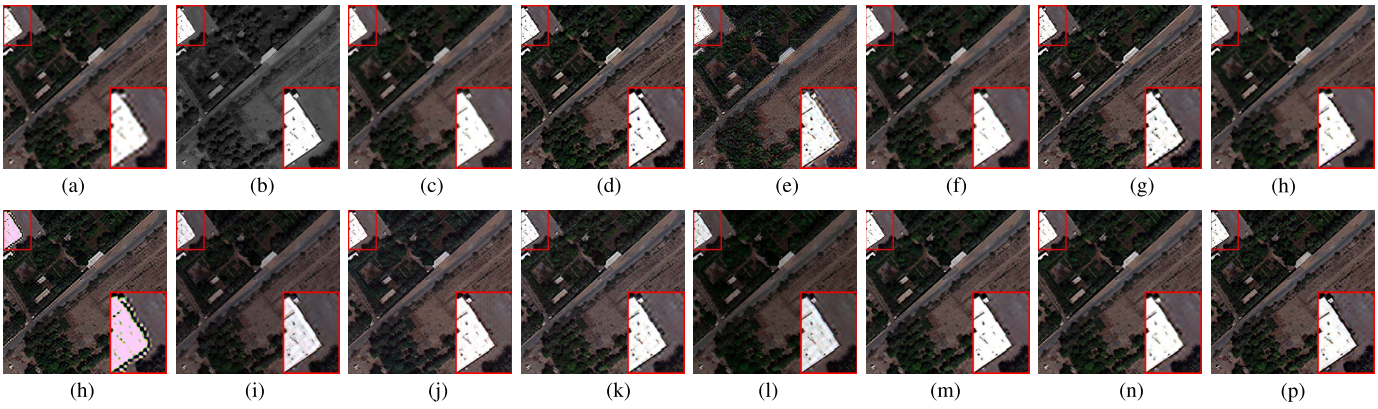


Fig. 9. Visual comparisons on the original scale data. As shown in the red rectangles, our model can produce sharper results compared with other methods, with fewer artifacts on the white roof. Please zoomed-in view for a better visualization. (a) LRMS. (b) PAN. (c) ATWT-M3. (d) AWLP. (e) BDSD. (f) PRACS. (g) Indusion. (h) PHLP. (i) SIRF. (j) PNN. (k) U-Nets. (l) RBDN. (m) ResNet. (n) ResNet + spec. (o) PanNet. (p) Our.

to the upsampled LRMS images are shown in Fig. 10. Since the output and upsampled LRMS images should have close spectral information, smooth regions should be close to zero, and only edges or structures should show, corresponding to information missing from the LRMS images.

We also adopt the strategy used in [25] to perform quantitative evaluations, i.e., we downsample the output HRMS images and compare them with the LRMS as ground truth. We also use QNR [56], which is composed of spectral distortion index D_λ and spatial distortion index D_s, as the reference-free measure.

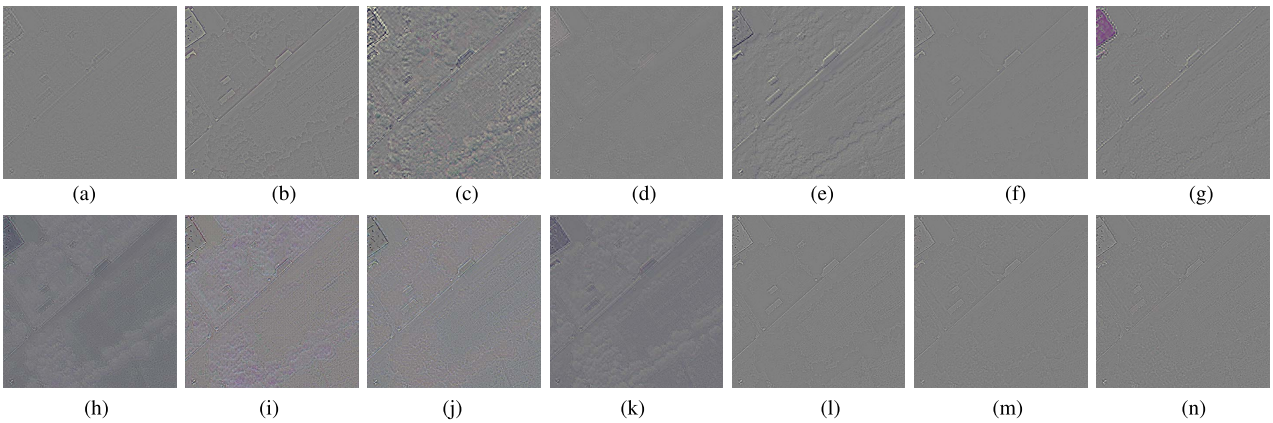


Fig. 10. Residuals between the HRMS image reconstructions and the upsampled LRMS image from Fig. 9. Our network focuses on reconstructing the high-frequency structure missing from LRMS. (a) ATWT-M3. (b) AWLP. (c) BDSD. (d) PRACS. (e) Indusion. (f) PHLP. (g) SIRF. (h) PNN. (i) U-Nets. (j) RBDN. (k) ResNet. (l) ResNet + spec. (m) PanNet (n) Our.

TABLE II

COMPARISON ON PARAMETER NUMBERS OF DEEP LEARNING-BASED METHODS

Methods	PNN [2]	U-Nets [42]	RNBN [43]	PanNet [4]	Our
Parameter #	$125K$	$7.7M$	$800K$	$80K$	$100K$

TABLE III

QUALITY METRICS OF DIFFERENT METHODS ON 200 SATELLITE IMAGES FROM WORLDVIEW3. THE BEST AND THE SECOND BEST RESULTS ARE BOLDFACED AND UNDERLINED, RESPECTIVELY

Methods	D_λ	D_s	QNR
ATWT-M3 [17]	0.076 ± 0.016	0.121 ± 0.021	0.812 ± 0.025
AWLP [16]	0.065 ± 0.026	0.108 ± 0.018	0.835 ± 0.037
BDSD [14]	0.079 ± 0.035	0.128 ± 0.034	0.803 ± 0.048
PRACS [13]	0.019 ± 0.006	0.103 ± 0.021	0.880 ± 0.021
Indusion [15]	0.055 ± 0.023	0.073 ± 0.018	0.876 ± 0.034
PHLP [23]	0.029 ± 0.020	0.077 ± 0.019	0.896 ± 0.035
SIRF [24], [25]	0.070 ± 0.027	0.088 ± 0.027	0.849 ± 0.047
PNN [2]	0.036 ± 0.008	0.093 ± 0.021	0.875 ± 0.022
U-Nets [42]	0.034 ± 0.016	0.066 ± 0.018	0.905 ± 0.015
RNBN [43]	0.026 ± 0.008	0.047 ± 0.012	0.928 ± 0.010
ResNet [26]	0.032 ± 0.012	0.105 ± 0.020	0.866 ± 0.022
ResNet+spectra-map	0.026 ± 0.009	0.085 ± 0.013	0.891 ± 0.017
PanNet [4]	0.023 ± 0.008	0.071 ± 0.013	0.908 ± 0.015
Our	**0.017 ± 0.006**	**0.054 ± 0.011**	**0.930 ± 0.013**
ideal value	0	0	1

These results are shown in Table III, where we can see the promising performance of our model again.

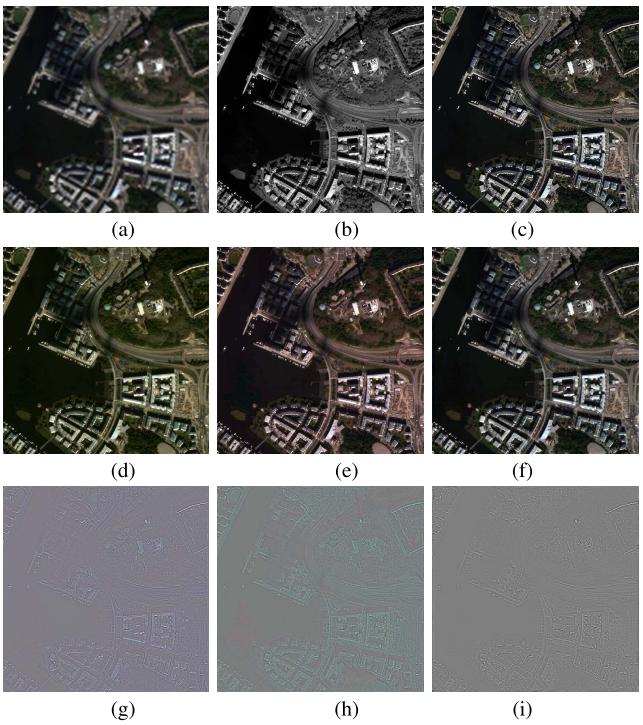


Fig. 11. Network generalization ability that tests on **WorldView2** [(e) and (f) are trained on **WorldView3**]. (a) LRMS. (b) PAN. (c) Ground Truth. (d) PNN-WV2 [2]. (e) PNN-WV3. (f) Our. (g) $|(c) - (d)|$. (h) $|(c) - (e)|$. (i) $|(c) - (f)|$.

C. Generalization to New Satellites

We have motivated our multiscale network as being more robust to differences across satellites. To show this, we compare our model with PNN on both the WorldView2 and WorldView3 satellite data sets. Specifically, two PNN-trained models are tested: one called PNN-WV2, which trained on WorldView2 data; the other called PNN-WV3 is trained on the same WorldView3 data set as our multiscale network.

We show one visual result in Fig. 11. We can see that PNN does not generalize well to new satellites, while our multiscale network can generalize well to WorldView2 being trained on WorldView3. In Fig. 11(d) and (e), PNN (-WV2 and -WV3, respectively) suffers from obvious spectral distortion, which our network is robust to new types of data. This affirms the ability of our model to leave modeling the spectral information to the spectra-mapping procedure and to allow the network to focus on structural information. On the other hand, PNN requires its network to model both spatial and spectral information.

We also considered how our model and PNN generalize to the IKONOS satellite data. Since IKONOS data contain four bands, the R, G, B, and infrared, bands from WorldView3 data are selected to train our model. PNN-IK and PNN-WV3 indicate two models that PNN trained on IKONOS data and WorldView3 data, respectively. As shown in Fig. 12, PNN-WV3 has a clear structure at the cost of spectral distortion. Though PNN-IK is directly trained on IKONOS, our method

TABLE IV
QUANTITATIVE COMPARISONS FOR DIFFERENT DILATED FACTORS (DF) ON 225 SATELLITE IMAGES WORLDVIEW3

DF	Q8	QAVE	SAM	ERGAS	SCC
1	0.921 ± 0.032	0.924 ± 0.036	4.117 ± 0.732	2.369 ± 0.392	0.932 ± 0.013
2	0.927 ± 0.030	0.929 ± 0.033	4.012 ± 0.705	2.327 ± 0.381	0.943 ± 0.011
3	0.931 ± 0.030	0.932 ± 0.032	3.874 ± 0.691	2.304 ± 0.377	0.952 ± 0.009
4 (default)	0.934 ± 0.029	0.935 ± 0.032	3.681 ± 0.684	2.292 ± 0.373	0.957 ± 0.009
5	0.936 ± 0.028	0.939 ± 0.031	3.624 ± 0.680	2.285 ± 0.372	0.961 ± 0.007
ideal value	1	1	0	0	1

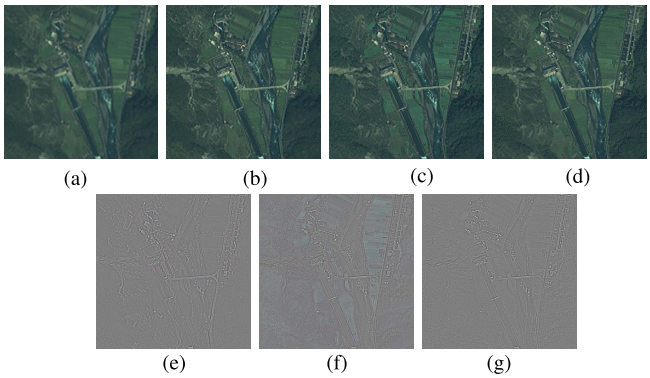


Fig. 12. Network generalization ability that tests on **IKONOS** [(c) and (d) are trained on **WorldView3**]. (a) LRMS. (b) PNN-IK [2]. (c) PNN-WV3. (d) Our. (e) $|(a) - (b)|$. (f) $|(a) - (c)|$. (g) $|(a) - (d)|$.

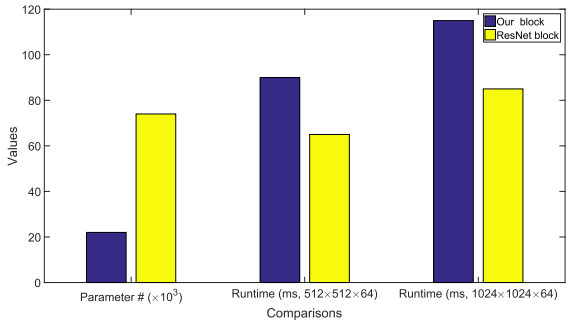


Fig. 13. Parameter numbers and runtime comparisons between ResNet block and our grouped dilated block. Note that the runtime is in milliseconds. Two different tensor sizes, i.e., 512 × 512 × 64 and 1024 × 1024 × 64, are used for testing.

still has clearer results. As can be seen in the second row of Fig. 12, i.e., the residual images, our method is able to simultaneously achieve spectral and spatial preservations. Specifically, compared with PNN, our result contains the less color difference in smooth regions and more clear structures around boundary regions. We see that using high-frequency parts to train our network can remove inconsistencies between different satellites.

D. Ablation Study

We provide ablation studies to explore the effect of each part of our model.

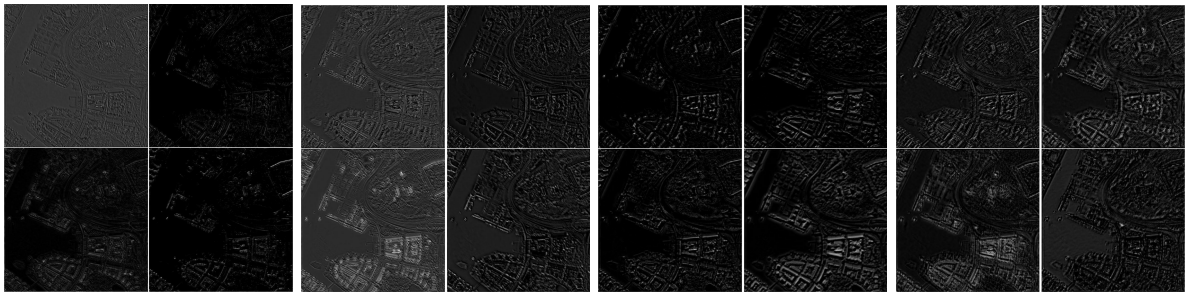
1) Effect of Grouped Dilated Blocks: Since introducing the grouped multiscale dilated blocks is the core of this article, we first test the effect of different dilated blocks by comparing it to the baseline network, which only contains normal convolutions, i.e., the dilated factor fixed as 1. Specifically, we test five different grouped dilated blocks by increasing the dilated factors from 2 to 5. In other words, the feature maps are divided into corresponding numbers of groups. Note that for a fair comparison, we have adjusted different dilated blocks to make them have close parameter numbers.

Quantitative results are shown in Table IV, and increasing dilated factors can generate better results. Adding dilated factors result in larger receptive fields, which have a greater advantage over the normal convolution. However, adding dilated factors eventually increase memory burdens and bring only limited improvement. Thus, to balance the tradeoff between performance and speed, we choose the maximum dilated factor equal to 4 as our default setting.

We also test the computational efficiency of the ResNet block and our grouped dilated block, i.e., the two network architectures, as shown in Fig. 6. As well known, under the convolutional neural network framework, the convolutional operations occupy the main running time. Due to the increase in the size of the convolutional kernel, directly adding dilated factors usually increases the running time. However, since we divide the feature map into four parallel groups, both the number of convolution kernels and the number of connections between adjacent features are reduced to a quarter of the original numbers. This gives our grouped dilated blocks more powerful multiscale representation abilities while maintaining similar computational efficiency as ResNet. As shown in Fig. 13, the parameter number of our block is reduced due to the grouping operation, while the computational runtime (in milliseconds) of each block is close after GPU acceleration.

To see which representations these modules have learned, we visualize some of the feature maps from different dilated convolutions in Fig. 14. Clearly, as the dilation factor increases, the corresponding feature maps contain larger scale structures and content, which is consistent with our intention to capture multiscale spatial patterns. Therefore, unlike the existing methods [42], [43] that perform multiscale representation in a layerwise manner, our grouped dilated block can extract multiscale features at a fine-grained level and increases receptive fields within one single-network layer. This brings a significant improvement in sharpening performance.

2) Effect of Removing BN: To demonstrate the effectiveness of removing BN as described in Section III-E, we conduct experiments by plugging BN into our deep model. The BN operation is deployed after each convolutional operation, which is the same as Fig. 6(a). We train the deep

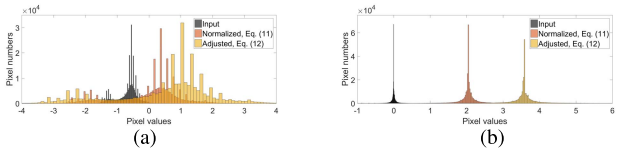


(a) (b) (c) (d)

Fig. 14. Examples of dilated convolutions. We only show four feature maps for each dilation factor for visualization. (a) $DF = 1$. (b) $DF = 2$. (c) $DF = 3$. (d) $DF = 4$.

TABLE V
QUANTITATIVE RESULTS OF REMOVING BN

	Worldview3 data		Worldview2 data	
Metrics	with BN	w/o BN	with BN	w/o BN
D_λ	0.019	0.017	0.035	0.029
D_s	0.058	0.054	0.084	0.067
QNR	0.924	0.930	0.884	0.906



(a) (b)

Fig. 15. Statistical histogram distributions of feature maps using BN. The parameters in (11) and (12) are learned from Worldview3 data and directly used on Worldview2 data. (a) Trained on Worldview3, tested on Worldview3. (b) Trained on Worldview3, tested on Worldview2.

model with BN on the Worldview3 data and test it on both Worldview3 and Worldview2 data at the original scale. The qualitative results are shown in Table V. It can be seen that on the testing data of Worldview3, the overall performances of the two models are very close. This is because both training data and testing data are collected from the same satellite, which meets the assumption of BN. While on the testing data of Worldview2, the performance of the model with BN is significantly reduced. This is because the distribution forms of the data collected by the two satellites are different. The distribution form learned by BN, i.e., the mean and standard deviation in (11), on one type of satellite cannot be directly used for another type of satellite.

To prove this viewpoint, in Fig. 15, we show an example of statistical histogram distributions of feature maps at the first layer. The parameters in (11) and (12) are learned from Worldview3 data. It is clear that in Fig. 15(a), if the training data and testing data have the same distribution form, i.e., both from Worldview3, the feature maps generated after the BN operation change in a relatively small range. On the contrary, using the learned parameters from Worldview3 to process Worldview2 data leads to an obvious change, as shown in Fig. 15(b). This may lead to subsequent calculations to be unstable, which degrades the performance. In addition, adding BN operations consumes more computing and storage resources. Therefore, we argue that adopting BN is not suitable for this specific remote sensing community. To improve the generalization ability and save the computing and storage budget, we remove BN from our model.

3) Effect of Hyperparameter Settings: We also tested the impact of kernel number and grouped dilated block number. Specifically, we first test the kernel numbers $K \in \{16, 32, 48, 64\}$ while fixing the grouped dilated block number as 4. Then, we test the grouped dilated block numbers $L \in \{1, 2, 4, 6\}$ while fixing the kernel number as 64. Quantitative results are shown in Table VI. It is clear that increasing kernels and blocks can generate higher performance. Adding grouped blocks results in larger modeling capacity and more nonlinear operations, which have a greater advantage over increasing the kernel numbers. However, increasing K and L eventually brings only limited improvement at the cost of storage and computation. We also adjust the PanNet to make its parameter numbers close to our proposed multiscale model. The quantitative results are shown in Table VII. It is clear that under different orders of magnitude, our proposed method consistently outperforms the PanNet, which further demonstrates the effectiveness of our model. To balance the tradeoff between performance and speed, we choose $K = 64$ and $L = 4$ as our default setting.

E. Convergence

We show the convergence on the training and testing data sets as a function of SGD iteration in Fig. 16. We focus on the four different network structures: ResNet, ResNet+spectramapping, PanNet, and our multiscale network. As can be seen in Fig. 16, our network has a significantly lower training and testing error than other network structures. This demonstrates that our multiscale network structure and high-pass training strategy are suitable for the specific pan-sharpening problem.

F. Hyperspectral Image Sharpening

Our proposed network can be directly extended to other applications. In this section, we test our model on HSI sharpening [62], which has attracted more and more attention in earth remote sensing tasks [63], e.g., object classification [64]–[66] and change detection [36]. This task aims to fuse an LR HSI (HSI) with an HR multispectral image to obtain an HR-HSI. We adopt the fusion framework reported in method [36] to evaluate our model. We have compared our network with four state-of-the-art methods, i.e., HYperspectral

TABLE VI

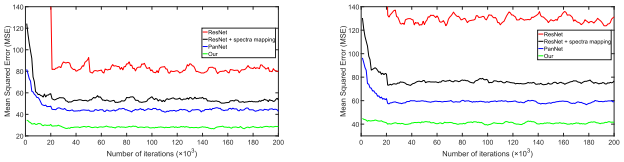
QUANTITATIVE COMPARISONS ON DIFFERENT NUMBERS OF KERNEL (K) AND BLOCK (L) ON 225 SATELLITE IMAGES WORLDVIEW3

Settings	Q8	QAVE	SAM	ERGAS	SCC
$K = 16, L = 4$	0.847 ± 0.039	0.849 ± 0.041	5.472 ± 1.253	3.858 ± 0.720	0.861 ± 0.017
$K = 32, L = 4$	0.864 ± 0.033	0.869 ± 0.037	4.574 ± 0.921	3.357 ± 0.562	0.882 ± 0.015
$K = 48, L = 4$	0.895 ± 0.032	0.901 ± 0.035	4.149 ± 0.871	2.946 ± 0.512	0.904 ± 0.012
$K = 64, L = 4$ (default)	0.934 ± 0.029	0.935 ± 0.032	3.681 ± 0.684	2.292 ± 0.373	0.957 ± 0.009
$K = 64, L = 1$	0.814 ± 0.034	0.817 ± 0.043	5.861 ± 1.147	4.264 ± 0.877	0.817 ± 0.019
$K = 64, L = 2$	0.881 ± 0.031	0.885 ± 0.035	4.243 ± 0.721	3.081 ± 0.504	0.894 ± 0.013
$K = 64, L = 6$	0.946 ± 0.027	0.947 ± 0.031	3.415 ± 0.624	2.133 ± 0.351	0.964 ± 0.008
ideal value	1	1	0	0	1

TABLE VII

QUANTITATIVE COMPARISONS WITH PANNET UNDER DIFFERENT PARAMETER NUMBERS ON 225 SATELLITE IMAGES WORLDVIEW3

Parameter #	Network	Q8	QAVE	SAM	ERGAS	SCC
$\approx 10K$	PanNet	0.823 ± 0.032	0.826 ± 0.036	5.974 ± 1.156	4.579 ± 0.848	0.846 ± 0.015
	Our	0.847 ± 0.039	0.849 ± 0.041	5.472 ± 1.253	3.858 ± 0.720	0.861 ± 0.017
$\approx 50K$	PanNet	0.875 ± 0.030	0.884 ± 0.032	4.472 ± 0.947	3.674 ± 0.537	0.887 ± 0.015
	Our	0.895 ± 0.032	0.901 ± 0.035	4.149 ± 0.871	3.081 ± 0.512	0.904 ± 0.012
$\approx 100K$	PanNet	0.927 ± 0.025	0.931 ± 0.023	3.927 ± 0.731	2.407 ± 0.403	0.948 ± 0.010
	Our	0.934 ± 0.029	0.935 ± 0.032	3.681 ± 0.684	2.292 ± 0.373	0.957 ± 0.009
ideal value		1	1	0	0	1



(a) (b)

Fig. 16. Convergence of different network structures. (a) Training error curves. (b) Testing error curves.

image super-resolution via SUbspace-based REgularization (HYSURE) [59], coupled spectral unmixing (CSU) [60], nonnegative-structured sparse representation (NSSR) [61] and deep hyperspectral image sharpening (DHSIS) [36]. Table VIII shows the quantitative results on two public data sets, i.e., CAVE data set [57] and Harvard data set [58]. As can be seen, our model has the best overall performance on the four metrics. This is because we adopt domain knowledge to, respectively, preserve spectral and spatial information, which is also the key role of HSI sharpening. This demonstrates that our multiscale network is a general model for different tasks.

In Fig. 17, we show the sharpening results from each data set for visual comparison. As shown in the red rectangles in output, the HR-HSIs, HYSURE, CSU, and NSSR methods have obvious spectral artifacts, while DHSIS has noticeable edge distortion. On the contrary, our model can simultaneously achieve spectral and spatial preservations. We also show the corresponding residual images using pseudocolors to reflect the differences between predicted HR-HSIs and ground truths. As shown in residual images, other compared methods contain various degradations, such as blurry details and ringing artifacts, especially in the marked regions. Our proposed deep multiscale detail network achieves the best performance in detail reconstruction and artifacts reduction. At the same time, our residual image is displayed in dark blue on the overall smooth area; that is, all difference values are close to 0. While other residual images more or less contain noticeable regions, indicating relatively large errors. This indicates that our method achieves better spectral preservation.

G. Discussion

Our approach belongs to data-driven methods and directly learns the relationship between inputs and desirable high-quality outputs. The physical models reflecting the satellite characterizes, and imaging processes are ignored. However, our deep model can be combined with handcrafted algorithms to take full advantage of both methods. For example, our network can be used as the regularization term, i.e., $J_3(\cdot)$ in (3), to implicitly express the complex prior from training data. In this way, both powerful representation ability of deep networks and prior physical models can be jointly exploited, which may further boost the performance. We leave this to our future work.

Our approach also belongs to supervised methods and predicts pixel values based on synthetic training data. On the other hand, the generative adversarial networks [67] are able to capture data distribution forms in an unsupervised manner. As described in the method [68], using adversarial learning generates realistic but not real results. We test the GANs' effect and show one visual result in Fig. 18. It is clear that using adversarial loss can generate sharper results with fake details, as shown in the enlarged regions. For the remote sensing community, accuracy is more important than perceptual qualities. Therefore, we do not introduce GANs to our method.

In method [69], the residual and high-frequency details are also applied to the image deraining. However, our method differs from this method in two ways. First, the purpose of these two tasks is different. The image deraining is a one-to-one mapping process in which the spatial resolution

TABLE VIII
QUANTITATIVE RESULTS ON CAVE [57] AND HARVARD [58] DATA SETS

	CAVE dataset [57]				Harvard dataset [58]			
Methods	PSNR	SAM	SSIM	ERGAS	PSNR	SAM	SSIM	ERGAS
HYSURE [59]	37.36	9.84	0.9451	2.01	43.88	4.20	0.9750	1.56
CSU [60]	41.66	6.29	0.9817	1.19	45.40	3.74	0.9816	1.36
NSSR [61]	44.70	4.31	0.9854	0.83	46.04	3.60	0.9816	1.25
DHSIS [36]	46.62	3.77	0.9912	0.64	46.42	**3.47**	**0.9827**	1.28
Ours	**47.62**	**3.45**	**0.9929**	**0.57**	**46.50**	3.48	**0.9827**	**1.09**
ideal value	$+\infty$	0	1	0	$+\infty$	0	1	0

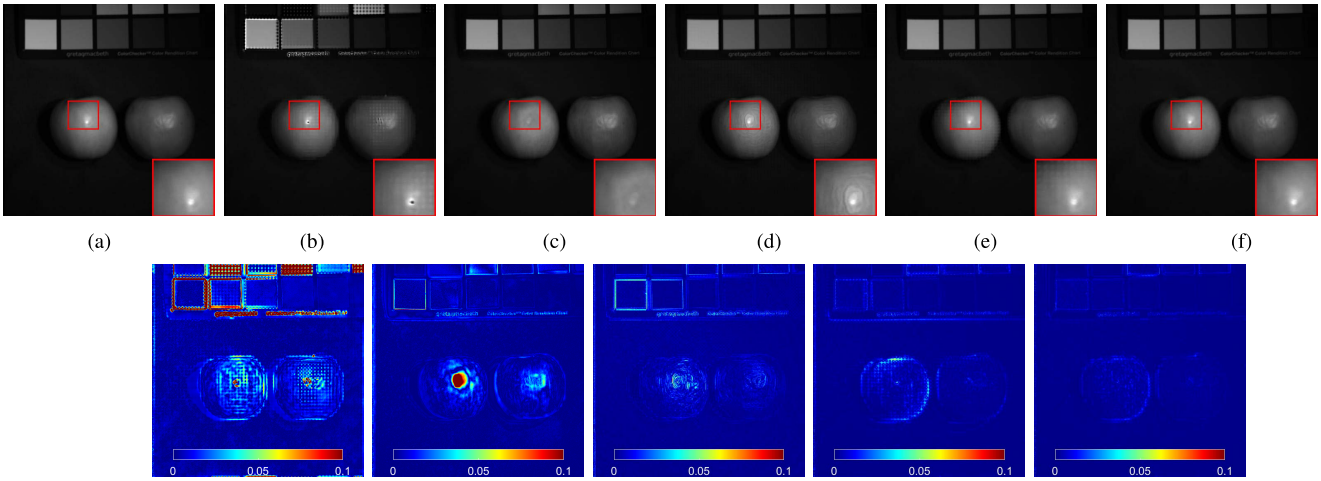


Fig. 17. One example (CAVE data set [57]) of HSI sharpening at the last band. All images and absolute residuals are normalized for a better visualization. (a) Ground truth. (b) HYSURE [59]. (c) CSU [60]. (d) NSSR [61]. (e) DHSIS [36]. (f) Our. (g) Residual | (a)−(b) |. (h) Residual | (a)−(c) |. (i) Residual | (a)−(d) |. (j) Residual | (a)−(e) |. (k) Residual | (a)−(f) |.

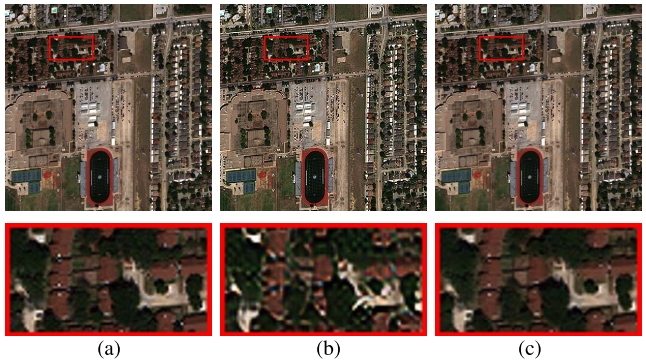


Fig. 18. Visual results of introducing GANs. The second row shows the enlarged areas in the first row. Using GANs generates sharper results with fake details. (a) Ground truth. (b) With GANs. (c) Without GANs.

and number of bands remain unchanged. This task aims to reconstruct a high-quality image from its low-quality observation, while the multiband spectral image sharpening is a fusion process. It receives two inputs with different spatial resolutions and band numbers and aims to output a high-quality image that contains all valuable spatial and spectral information. Second, in addition to simplifying the learning process, the two approaches have different motivations for applying residual and high-frequency details. For the image deraining, both rain streaks and objects details belong to high-frequency parts. Using residual and high-frequency details helps the network effectively extract rain streaks patterns. For the multiband spectral image sharpening, using residual and high-frequency details aims to achieve the fusion and preservation of spatial information, which is one of the goals of this task.

V. CONCLUSION

In this article, we propose a deep CNN-based approach to sharpen multiband spectral images by combining deep learning technology with domain-specific knowledge. Based on the analysis of the pansharpening problem, we designed our network by taking into consideration the two goals—spectral and spatial preservations. To preserve spectral information, we introduce a concept of spectra mapping that directly adds upsampled LRMS images to the squared objective term to allow the network to focus on high-frequency parts. To preserve spatial information, by using a multiscale fashion with dilated convolutions, we train our network on the high-frequency components of the PAN and upsampled LRMS images. This multiscale structure allows us to fuse spatial information at different scales to further improve the reconstruction accuracy. Compared with the state-of-the-art methods, our method achieves better image fusion and generalization to new satellites. We also test our model on the HSI sharpening to show the potential value of our network for different tasks.

REFERENCES

[1] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[2] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[4] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5449–5457.

[5] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Jan. 2002.

[6] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

[7] 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.

[8] G. Vivone *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[9] 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.

[10] W. J. Carper, T. M. Lillesand, and R. W. Kiefer, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 457–467, 1990.

[11] P. S. Chavez, Jr., and A. Y. Kwarteng, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, 1989.

[12] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, Aug. 1987.

[13] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[14] 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.

[15] M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert, "Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 98–102, Jan. 2008.

[16] X. Otazu, M. González-Audícana, O. Fors, and J. Núñez, "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.

[17] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 1, pp. 49–61, Jan. 2000.

[18] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[19] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 43–58, Aug. 2006.

[20] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2822–2834, Jul. 2013.

[21] X. Ding, Y. Jiang, Y. Huang, and J. Paisley, "Pan-sharpening with a Bayesian nonparametric dictionary learning model," in *Proc. AISTATS*, 2014, pp. 176–184.

[22] H. A. Aly and G. Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2596–2608, Jun. 2014.

[23] Y. Jiang, X. Ding, D. Zeng, Y. Huang, and J. Paisley, "Pan-sharpening with a hyper-Laplacian penalty," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 540–548.

[24] C. Chen, Y. Li, W. Liu, and J. Huang, "SIRF: Simultaneous satellite image registration and fusion in a unified framework," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4213–4224, Nov. 2015.

[25] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2760–2765.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[28] M. P. Eckstein, H. Cecotti, and B. Giesbrecht, "Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2030–2042, Feb. 2014.

[29] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.

[30] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[31] X. Hu, G. Feng, S. Duan, and L. Liu, "A memristive multilayer cellular neural network with applications to image processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1889–1901, Aug. 2017.

[32] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 21, 2020, doi: 10.1109/TPAMI.2020.2968521.

[33] J. Xie, L. Xu, E. Chen, J. Xie, and L. Xu, "Image denoising and inpainting with deep neural networks," in *Proc. NIPS*, 2012, pp. 341–349.

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[35] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2944–2956, Jun. 2017.

[36] 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.

[37] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by MS/HS fusion net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1585–1594.

[38] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[39] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[40] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, 2018, pp. 286–301.

[41] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 576–584.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[43] V. Santhanam, V. I. Morariu, and L. S. Davis, "Generalized deep image to image regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5609–5619.

[44] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5344–5353.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[46] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, 2016, pp. 1–13.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[48] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

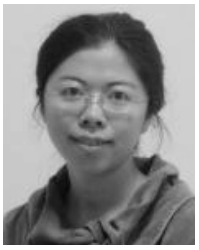
[49] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 675–678.
[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[51] L. Wald, *Data fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses des MINES, 2002.
[52] R. H. Yuhas, A. F. H. 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. 1–3.
[53] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Aug. 2002.
[54] 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, Oct. 2004.
[55] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.
[56] 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.
[57] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.
[58] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.
[59] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.
[60] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.
[61] W. Dong *et al.*, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.
[62] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019, doi: 10.1109/TNNLS.2018.2885616.
[63] Y. Gao, X. Wang, Y. Cheng, and Z. J. Wang, "Dimensionality reduction for hyperspectral data based on class-aware tensor neighborhood graph and patch alignment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1582–1593, Aug. 2015.
[64] N. Akhtar and A. Mian, "Nonparametric coupled Bayesian dictionary and classifier learning for hyperspectral classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4038–4050, Sep. 2018.
[65] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016.
[66] P. Zhong and R. Wang, "Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1319–1334, Jul. 2014.
[67] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NuerIPS*, 2014, pp. 2672–2680.
[68] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[69] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3855–3863.

Xueyang Fu (Member, IEEE) received the Ph.D. degree in signal and information processing from Xiamen University, Xiamen, China, in 2018.

He was a Visiting Scholar with Columbia University, New York, NY, USA, sponsored by the China Scholarship Council, from 2016 to 2017. He is currently an Associate Researcher with the Department of Automation, University of Science and Technology of China, Hefei, China. His research interests include machine learning and image processing.

Wu Wang received the B.S. degree from the Hefei University of Technology, Hefei, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Informatics, Xiamen University, Xiamen, China.

His research interests mainly focus on machine learning and image processing and analysis.



Yue Huang received the B.S. degree from Xiamen University, Xiamen, China, in 2005, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2010.

She was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2015 to 2016. She is currently an Associate Professor with the School of Informatics, Xiamen University. Her main research interests include machine learning and image processing.

Xinghao Ding (Member, IEEE) was born in Hefei, China, in 1977. He received the B.S. and Ph.D. degrees from the Department of Precision Instruments, Hefei University of Technology, Hefei, in 1998 and 2003, respectively.

He was a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2009 to 2011. Since 2011, he has been a Professor with the School of Informatics, Xiamen University, Xiamen, China. His main research interests include machine learning, representation learning, medical image analysis, and computer vision.

John Paisley (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Duke University, Durham, NC, USA, in 2004, 2007, and 2010, respectively.

He was a Post-Doctoral Researcher with the Department of Computer Science, University of California at Berkeley, Berkeley, CA, USA, and the Department of Computer Science, Princeton University, Princeton, NJ, USA. He is currently an Associate Professor with the Department of Electrical Engineering, Columbia University, New York, NY, USA, where he is also a member of the Data Science Institute. His current research interest is machine learning with an emphasis on models and inference techniques for text and image processing applications.