SDPNet: A Deep Network for Pan-Sharpening With Enhanced Information Representation

Han Xu, Jiayi Ma, Zhenfeng Shao, *Member, IEEE*, Hao Zhang, Junjun Jiang, *Member, IEEE*, and Xiaojie Guo, *Senior Member, IEEE*

***Abstract*— In this article, we propose a surface- and deep-level constraint-based pan-sharpening network, termed SDPNet, to address the pan-sharpening problem. Focusing on the two primary goals of pan-sharpening, i.e., spatial and spectral information preservations, we first design two encoder–decoder networks to extract deep-level features from two types of source images, in addition to surface-level characteristics, as the enhanced information representation. The unique feature maps that characterize the unique information in source images can be obtained through the deep-level feature extraction. We further design a pan-sharpening network with densely connected blocks to strengthen feature propagation and reduce parameter number, where the unique feature maps are utilized to efficiently constrain the similarity between the pan-sharpened result and the ground truth, thus avoiding information distortion. Both qualitative and quantitative comparisons on the reduced-resolution and full-resolution source images demonstrate the advantages of our method over state-of-the-art methods. Our code is publicly available at https://github.com/hanna-xu/SDPNet.**

***Index Terms*— Encoder–decoder, feature extraction, image fusion, pan-sharpening.**

I. Introduction

With the launch of several optical earth observation satellites, many data captured by them can be used for various tasks, such as environment monitoring, geography, agriculture, and land survey. However, due to the limitation of physical techniques, it is difficult for satellites to combine high spatial and high spectral resolutions simultaneously. The captured data are usually in two modalities: the high-resolution panchromatic (PAN) image with low spectral resolution and the low-resolution multispectral (LRMS) image with high spectral resolution. Depending on the complementarity between these two modalities, a high-resolution multispectral (HRMS) image can be produced with both high spatial and spectral resolutions by fusing PAN and LRMS images, as shown in Fig. 1. This is the goal of pan-sharpening to break through the technical limitation.

In recent years, many different kinds of traditional methods have been put forward to solve this problem. According to the corresponding theories, they can be divided into the following categories: 1) methods based on multiscale decomposition, including pyramid [2], [3], contourlet [4], nonnegative matrix factorization [5], and principle component analysis [6]; 2) methods based on component substitution, e.g., the improved component substitution pansharpening through multivariate regression [7], adaptive component substitution using partial replacement [8], and a novel component substitution framework based on image matting [9]; 3) methods based on model optimization, such as the sparsity regularization-based method [10] and the optimum algorithm in the minimum mean-square-error sense [11]; and 4) hybrid methods, e.g., the HCM algorithm [12], which integrates the hybrid color and plug-and-play algorithms. However, considering the different spectral responses of various sensors and the complexity of ground objects, it is difficult to formulate the relationship between source images and the HRMS images in traditional ways.

Over the past few years, benefiting from the wide application of deep learning, scholars have attempted to draw support from the high nonlinearity of convolutional neural networks (CNNs) to solve the pan-sharpening problem [13]. A well-known CNN-based method is PNN [14]. It modifies the three-layer architecture of a super-resolution method SRCNN [15] by applying some specific knowledge in remote sensing. Also, based on a three-layer CNN, Zhong *et al.* [16] presented a hybrid pan-sharpening method. It employs CNNs to enhance the spatial resolution of multispectral (MS) images; then, the Gram–Schmidt transformation is utilized to fuse the enhanced MS and PAN images. Moreover, based on the domain-knowledge, Yang *et al.* [17] proposed PanNet by directly propagating the upsampled LRMS image to the output of the network to preserve the spectral information and training the network in the high-pass filtering domain to preserve the spatial structure. Furthermore, by improving the architecture, Wei *et al.* [18] put forward DRPNN by applying a deeper network to learn the residuals (spatial details) between the LRMS image and the ground truth. Besides, based on a

Manuscript received July 28, 2020; revised August 21, 2020; accepted September 3, 2020. Date of publication September 18, 2020; date of current version April 22, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61773295, Grant 41890820, Grant 61971165, and Grant 61772512; and in part by the Natural Science Foundation of Hubei Province under Grant 2019CFA037. *(Corresponding author: Zhenfeng Shao.)*

Han Xu, Jiayi Ma, and Hao Zhang are with Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: xu_han@whu.edu.cn; jyma2010@gmail.com; zhpersonalbox@gmail.com).

Zhenfeng Shao is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: shaozhenfeng@whu.edu.cn).

Junjun Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: jiangjunjun@hit.edu.cn).

Xiaojie Guo is with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: xj.max.guo@gmail.com).

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/TGRS.2020.3022482

Fig. 1. Illustration of the pan-sharpening problem. From left to right: LRMS image, PAN image, pan-sharpened results of LGC [1] and our proposed SDPNet, and the ground truth (HRMS image).

target-adaptive usage modality, Scarpa *et al.* [19] proposed TACNN that ensures a very good performance in the presence of a mismatch with respect to the training set and even across different sensors. In addition to the abovementioned series of works, there are also some works based on other models. For instance, based on generative adversarial networks (GANs) [20], [21], Liu *et al.* [22] proposed PSGAN to increase the similarity from the perspective of probability distribution. By combining the idea of an autoencoder with GAN, Shao *et al.* [23] proposed RED-cGAN by adopting the residual encoder–decoder module to extract the multiscale features and applying a conditional discriminator to encourage that the estimated MS images share the same distribution as that of the referenced HRMS images. By contrast, Ma *et al.* [24] proposed an unsupervised method, termed Pan-GAN, where the generator separately establishes the adversarial games with the spectral discriminator and the spatial discriminator, so as to preserve the rich spectral information of MS images and the spatial information of PAN images. As a combination of traditional algorithm and deep learning, Zhang *et al.* [25] proposed an efficient bidirectional pyramid network to process MS and PAN images in two separate branches level by level. Recently, according to the proportional maintenance of gradient and intensity, a general fusion framework named PMGI was proposed in [26], which can be applied to solve the pan-sharpening problem.

Although the existing methods can obtain good results, there are still some problems to be solved. On the one hand, many methods train the network by minimizing the Euclidean distance between the generated HRMS image and the ground truth, leading to relatively blurred results. To solve this problem, in some methods, the spectral or spatial information can be further preserved by additional operations, e.g., training in the high-pass filtering domain or learning the residuals. However, these operations are typically made in a manual way and still suffer from limitations, such as the appropriateness of feature or domain selection. On the other hand, it is difficult to define the spatial/spectral information comprehensively. Such information can be simply defined as surface-level characteristics, e.g., the high-frequency component and the pixel intensity. However, these features, in turn, are not enough to represent the spatial/spectral information in satellite images completely. Moreover, the spatial/spectral information does not merely exist in one type of satellite image. Instead, both PAN and LRMS images contain these two types of information simultaneously [27]. Therefore, the predefined characteristics may not reflect the unique information contained in one type of source images compared with the other one.

The abovementioned challenges motivate us to propose a new pan-sharpening network based on both surface- and deep-level constraints, i.e., SDPNet. The overall procedure of the proposed SPNet is shown in Fig. 2, which consists of three stages. In the first stage, we train M2PNet and P2MSNet to learn the transformations between the MS image and the corresponding PAN image of the same spatial resolution. In the second stage, spatial and spectral encoders and decoders are learned to extract the feature maps (including unique feature maps and common feature maps) and reconstruct the original images. In the last stage, we use the pretrained spatial and spectral encoders to perform the deep-level constraint. Based on both the deep- and surface-level constraints, PNet is trained to generate the pan-sharpened results.

The characteristics and contributions of our model are summarized as follows.

1) We design two encoder–decoder networks to extract deep-level features in addition to surface-level characteristics for enhanced spatial and spectral information representations, respectively. The deep-level features allow us to further minimize the difference between the pan-sharpened result and the ground truth.
2) Instead of preserving the manually predefined spatial-/spectral-related characteristics, we focus on preserving the unique features in each type of source images extracted by the corresponding encoder to improve the effectiveness of constraints. These unique features play a role of spatial-/spectral-related features for better information preservation.
3) Based on the two encoder–decoder networks, a pan-sharpening network is further designed by introducing densely connected blocks to strengthen feature propagation while reducing the number of parameters. Both qualitative and quantitative results confirm that the proposed SDPNet can outperform state-of-the-art methods with less spatial and spectral distortions.

II. Proposed Method

We denote the LRMS image as $\mathbf{M}$ of size $W \times H \times B$ and the high-resolution PAN image as $\mathbf{P}$ of size $rW \times rH \times 1$, where W and H are the width and height of the LRMS image,

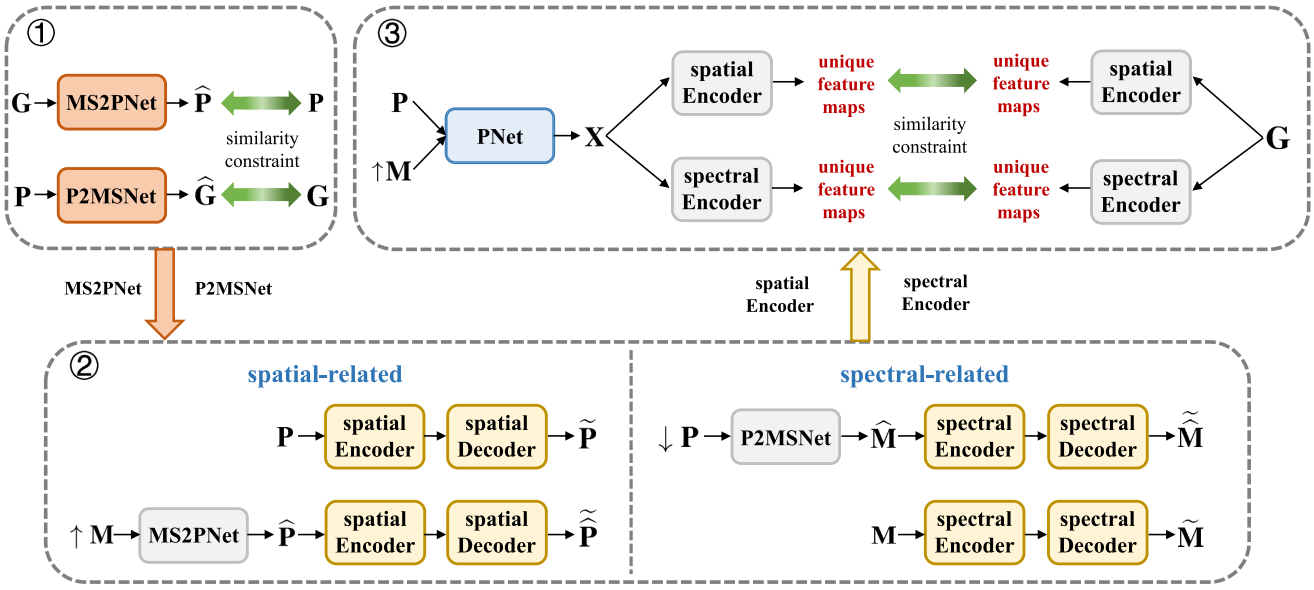

Fig. 2. Overall framework of the proposed SDPNet. **P**, **M**, **G**, and **X** denote the PAN, LRMS, HRMS (ground truth), and pan-sharpened images, respectively. $\widehat{\mathbf{I}}$ means the fake image transformed from the other type of image, and $\widetilde{\mathbf{I}}$ denotes the reconstructed image of an image **I**. ↑ and ↓ denote the upsampling and downsampling operations. The networks in gray indicate that the networks have been trained in the previous stage, and their parameters are fixed with no need for training. In the testing phase, only the trained PNet is needed to generate the pan-sharpened results.

respectively. B is the number of bands, and r is the spatial resolution ratio between **P** and **M**. The HRMS image is the ground-truth data for supervised learning, which is denoted as **G** of size $rW \times rH \times B$. Thus, the purpose of our work is to learn a model f_{p}, of which the output $\mathbf{X} = f_{\mathrm{p}}(\mathbf{P}, \mathbf{M})$ can be taken as the approximation of **G**.

A. Surface-Level Characteristics

Since the two priorities of pan-sharpening are the preservation of spatial and spectral information, we mainly constrain the similarity between **X** and **G** from these two aspects. Usually, the spatial information is supposed to mainly exist in spatial structures, while the spectral information is mainly characterized by the pixel intensity of each band in the MS image. Thus, by maximizing spatial and spectral similarities between **X** and **G**, the problem can be formulated as

$$f_{\mathrm{p}} = \arg\min_{\theta_{\mathrm{p}}} \sum_{b=1}^{B} 1 - \mathrm{SSIM}(\mathbf{X}_b, \mathbf{G}_b) + \lambda \|\mathbf{X}_b - \mathbf{G}_b\|_F^2 \quad (1)$$

where $1 - \mathrm{SSIM}(\mathbf{X}_b, \mathbf{G}_b)$ is the constraint of the spatial structure information for spatial preservation because SSIM is the structural similarity index measure [28] focusing on light, contrast, and structural information. $\|\mathbf{X}_b - \mathbf{G}_b\|_F^2$ denotes the constraint of the pixel intensity for spectral preservation with $\|\cdot\|_F$ denoting the Frobenius norm. $\mathbf{X}_b$ denotes the bth band of the generated HRMS image with B bands. Similarly, $\mathbf{G}_b$ denotes the bth band of the ground truth. θ_{p} represents the parameters to be optimized in the model f_{p}. λ is a positive number to control the tradeoff. Thus, these two terms can be employed to maximize the spatial and spectral similarities, respectively.

B. Deep-Level Features

Besides the abovementioned surface-level characteristics, there are some extra features that are outside the constraint terms of SSIM and the Frobenius norm. For instance, SSIM does not handle large displacements, nor assesses geometric deformations [29]. It will become unstable when the variance or luminance of the reference image is low [30]. For the Frobenius norm, all plausible outputs will be averaged. Thus, such a constraint may generate relatively blurred pan-sharpened results, leading to both spatial and spectral distortions. Certainly, the extra features are more than these. To make up for the insensitivity of surface-level metrics, drawing support from the high nonlinearity and strong learning ability of CNNs for capturing features, we perform a compensatory similarity constraint on deep-level features extracted by encoder–decoder networks.

1) Spatial-Related Features: The high-quality spatial structures are the unique information contained in **P** while not available in the bands of **M**. In order to extract the unique spatial-related features, we assume them as the most different features in **P** from **M**. To this end, a pseudo-PAN image $\widehat{\mathbf{P}}$ can be constructed by the LRMS image as

$$\widehat{\mathbf{P}} = f_{\mathrm{MS2P}}(\uparrow \mathbf{M}) \quad (2)$$

where f_{MS2P} is a function that can transform an MS image to a PAN image with the same resolution.

To learn the mapping relationship from an MS image to a PAN image, i.e., f_{MS2P}, we use a network, termed MS2PNet, to learning the mapping between different channels. As the mapping relationship has nothing to do with the spatial resolution difference, the ground-truth image **G** with high spatial resolution rather than the LRMS image and the corresponding PAN image are used as the training data to train the MS2PNet.

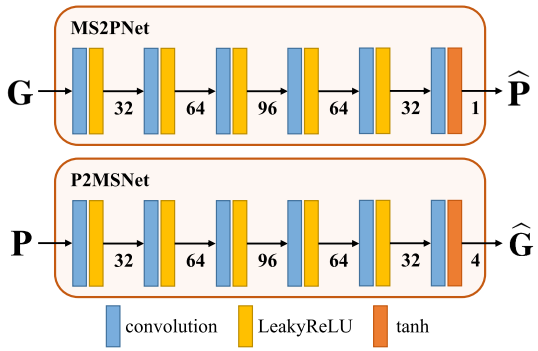

Fig. 3. Network architecture of the MS2PNet and P2MSNet.

To learn f_{MS2P}, the parameters in MS2PNet are optimized with the AdamOptimizer to maximize the similarity between $\mathbf{P}$ and $f_{\mathrm{MS2P}}(\mathbf{G})$ as

$$f_{\mathrm{MS2P}} = \arg\min_{\theta_{\mathrm{MS2P}}} \; 1 - \mathrm{SSIM}(\mathbf{P}, f_{\mathrm{MS2P}}(\mathbf{G})) + \lambda \|\mathbf{P} - f_{\mathrm{MS2P}}(\mathbf{G})\|_F^2 \quad (3)$$

where θ_{MS2P} denotes the parameters in the MS2PNet. By minimizing the loss function defined in (3), we can learn the optimal solution of f_{MS2P}. The network architecture of MS2PNet is shown in Fig. 3. The numbers of feature maps are shown after the activation function, including LeakyReLU and tanh. The kernel size is 3×3 and the stride is 1.

With the prelearned f_{MS2P}, we can generate a pseudo-PAN image from the LRMS image as in (2). Although there are some similar components, e.g., the similar pixel intensity distribution, contained in both $\mathbf{P}$ and $\widehat{\mathbf{P}}$ because they are the representations of the same scene, the high-quality structural information is usually only available in $\mathbf{P}$, as shown in Fig. 4. Although there are some structural differences in the common feature maps of $\mathbf{P}$ and $\widehat{\mathbf{P}}$ in Fig. 4, the differences of the structural details in unique feature maps are more obvious.

After being passed through the pretrained siamese networks (the training process of these networks will be described later), i.e., networks with the same architecture and parameters, some feature maps with larger differences can be regarded as unique feature maps of $\mathbf{P}$ or $\widehat{\mathbf{P}}$. According to the obvious difference between $\mathbf{P}$ and $\widehat{\mathbf{P}}$, feature maps extracted in these channels can be regarded as spatial-related features, as shown in the unique feature maps in Fig. 4. We define them as $\phi_{\mathrm{spat}}^{\mathbf{I},1}$, $\phi_{\mathrm{spat}}^{\mathbf{I},2}, \ldots, \phi_{\mathrm{spat}}^{\mathbf{I},N}$, where N is the number of unique feature maps and $\mathbf{I}$ is the input of this network, which can be specifically set as $\mathbf{P}$ or $\widehat{\mathbf{P}}$. Comparatively, those feature maps with smaller differences contain more information that is available in both $\mathbf{P}$ and $\widehat{\mathbf{P}}$, as shown in common feature maps in Fig. 4. Therefore, this encoder–decoder network can be regarded as a spatial encoder–decoder network, of which the specific architecture is shown in Fig. 5 with the numbers representing the channels of output feature maps.

For the training phase of this network, instead of sequential training, both $\mathbf{P}$ and $\widehat{\mathbf{P}}$ are used as the training data for jointly training. The spatial encoder–decoder network is trained by maximizing the similarity between $\mathbf{P}$ and the reconstructed PAN image $\widetilde{\mathbf{P}}$ and the similarity between $\widehat{\mathbf{P}}$ and the reconstructed pseudo-PAN image $\widetilde{\widehat{\mathbf{P}}}$. The measurement of similarity is the same as that defined in (1) except that B is set as 1 and $(\mathbf{X}, \mathbf{G})$ is replaced by $(\mathbf{P}, \widetilde{\mathbf{P}})$ when the input of the network is $\mathbf{P}$. In another case, when the input of the network is $\widehat{\mathbf{P}}$, $(\mathbf{X}, \mathbf{G})$ is replaced by $(\widehat{\mathbf{P}}, \widetilde{\widehat{\mathbf{P}}})$. As for the experiment settings, the network is trained with ten epochs with a batch size of 10. The parameters are denoted as θ_{spat} and updated by AdamOptimizer with a learning rate 0.002 and exponential decay.

2) Spectral-Related Features: Similarly, the B-band spectral information is the unique information contained in $\mathbf{M}$ compared with $\mathbf{P}$. To extract the unique spectral-related features, we construct a pseudo-LRMS image $\widehat{\mathbf{M}}$ by using $\mathbf{P}$, which can be defined as

$$\widehat{\mathbf{M}} = f_{\mathrm{P2MS}}(\downarrow \mathbf{P}) \quad (4)$$

where $\downarrow$ denotes the downsampling operation.

Analogously, to learn the mapping relationship from a PAN image to an MS image, i.e., f_{P2MS}, we design a network, termed P2MSNet. The training data are still the PAN image and the corresponding ground-truth image $\mathbf{G}$ with the same spatial resolution rather than the LRMS image with lower spatial resolution. To learn f_{P2MS}, the parameters in P2MSNet θ_{P2MS} are optimized with the AdamOptimizer to maximize the similarity between $\mathbf{G}$ and $f_{\mathrm{P2MS}}(\mathbf{P})$ as

$$f_{\mathrm{P2MS}} = \arg\min_{\theta_{\mathrm{P2MS}}} \sum_{b=1}^{B} 1 - \mathrm{SSIM}(\mathbf{G}_b, f_{\mathrm{P2MS}}(\mathbf{P})_b) + \lambda \|\mathbf{G}_b - f_{\mathrm{P2MS}}(\mathbf{P})_b\|_F^2. \quad (5)$$

By solving the problem defined in (5), the optimal solution of f_{P2MS} can be learned to perform the transformation defined in (4). The network architecture of P2MSNet is shown in Fig. 3, which is similar to that of the MS2PNet except that the input is of one channel and the output is of four channels.

As can be seen from the unique and common feature maps shown in Fig. 6, which are extracted from $\mathbf{M}$ and $\widehat{\mathbf{M}}$ by the pretrained encoder–decoder network shown in Fig. 7, the common feature maps of them share both the similar pixel intensity distribution and texture details. However, the pixel intensity distribution of their unique features varies greatly. By comparison, the unique feature maps of $\mathbf{M}$ exhibit similar structures with $\widehat{\mathbf{M}}$ but more abundant pixel intensity distribution. The unique and abundant pixel intensity can be regarded as the representation of spectral information. Therefore, we define these unique features as spectral-related features: $\phi_{\mathrm{spec}}^{\mathbf{I},1}, \phi_{\mathrm{spec}}^{\mathbf{I},2}, \ldots, \phi_{\mathrm{spec}}^{\mathbf{I},N}$, where $\mathbf{I}$ can be specifically set as $\mathbf{M}$ or $\widehat{\mathbf{M}}$. Therefore, the encoder–decoder network in Fig. 7 can be considered as a spectral encoder–decoder network.

Both $\mathbf{M}$ and $\widehat{\mathbf{M}}$ are used for jointly training the spectral encoder–decoder network. It is trained by maximizing the similarity between $\mathbf{M}$ and the reconstructed LRMS image $\widetilde{\mathbf{M}}$ and the similarity between the pseudo-LRMS image $\widehat{\mathbf{M}}$ and its reconstruction $\widetilde{\widehat{\mathbf{M}}}$. The loss function is defined in the same way as (1) except that $(\mathbf{X}, \mathbf{G})$ are replaced by $(\mathbf{M}, \widetilde{\mathbf{M}})$ or $(\widehat{\mathbf{M}}, \widetilde{\widehat{\mathbf{M}}})$. B is specifically set as 4. The network is also

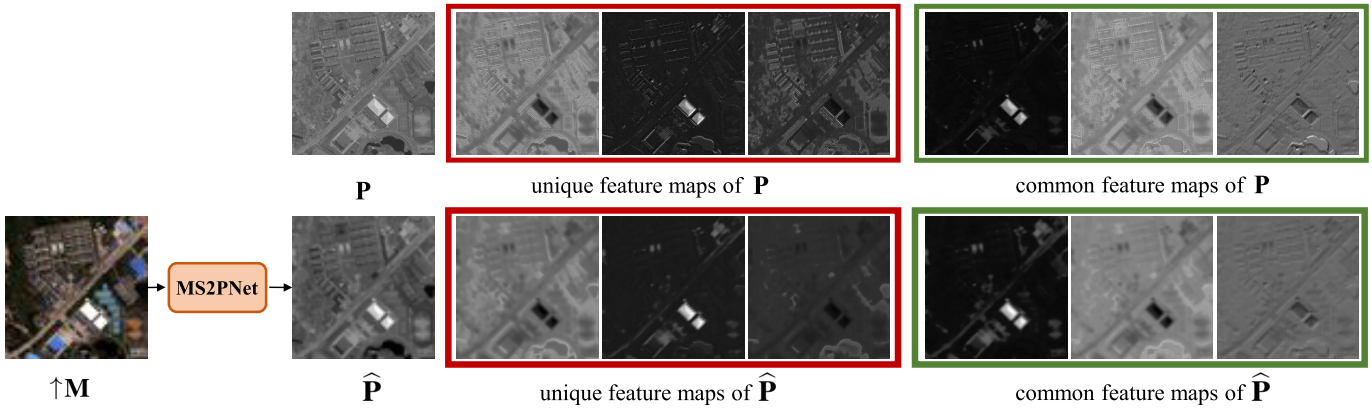

Fig. 4. Illustration of $\mathbf{P}$, $\widehat{\mathbf{P}}$ and some unique/common feature maps extracted from them by the spatial encoder–decoder network.

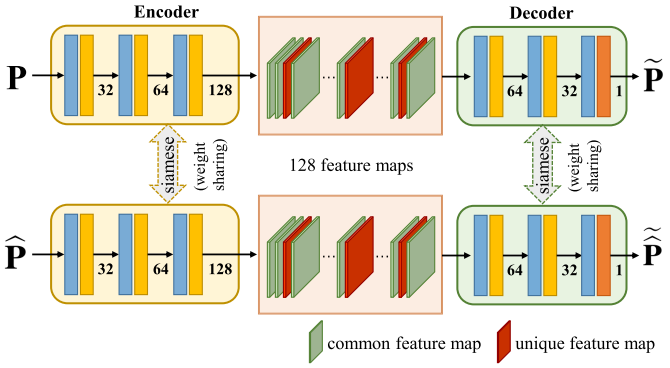

Fig. 5. Network architecture of the spatial encoder–decoder.

trained with a batch size of 10. The number of epochs is set as 10. The parameters are denoted as θ_{spec} and updated by AdamOptimizer with a learning rate 0.002 and exponential decay.

C. PNet

By comprehensively taking account of the similarity between the generated data $\mathbf{X}$ and the ground-truth data $\mathbf{G}$ in both surface-level characteristics and deep-level features, the pan-sharpening problem can be modified as

$$f_{\text{p}} = \arg\min_{\theta_{\text{p}}} \sum_{b=1}^{B} \underbrace{1 - \text{SSIM}(\mathbf{X}_b, \mathbf{G}_b) + \lambda \|\mathbf{X}_b - \mathbf{G}_b\|_F^2}_{\text{surface-level constraint}} + \frac{\gamma}{N} \sum_{n=1}^{N} \underbrace{\|\phi_{\text{spat}}^{\mathbf{X},n} - \phi_{\text{spat}}^{\mathbf{G},n}\|_1 + \alpha \|\phi_{\text{spec}}^{\mathbf{X},n} - \phi_{\text{spec}}^{\mathbf{G},n}\|_1}_{\text{deep-level constraint}}. \quad (6)$$

To solve the model f_{p} in (6), we design a network, termed PNet, with the parameters to be updated denoted as θ_{p}. The input of PNet is the concatenation of $\mathbf{P}$ and $\uparrow \mathbf{M}$ in the channel dimension. The architecture of PNet is shown in Fig. 8. There are eight layers where each layer consists of a convolutional layer and the following activation function. To train PNet more efficiently and improve the information flow, inspired by the densely connected blocks in [31], we build short connections in the second to fifth layers. More concretely, direct connections are built between layers close to the input and those close to the output in a feedforward fashion. These connections can alleviate the problems of vanishing gradients and strengthen feature propagation to improve network performance while reducing the number of parameters [32]. The subsequent three layers gradually reduce the number of feature maps until generating $\mathbf{X}$.

For the specific settings of each layer, the numbers of feature maps are shown after the activation function. The kernel size of the convolutional layer is set as 3×3, and the stride is set as 1. We employ reflection padding to reduce boundary artifacts. The activation function of the first seven layers is LeakyReLU with the slope set as 0.2 except that the activation function of the last layer is tanh.

III. Experiments and Result Analysis

We provide the details of the data set and training phase. Both the visual inspection and quantitative comparison are performed to validate the effectiveness of our method. Ablation experiments are conducted to verify the contribution of each component of our method.

A. Data Set and Training Details

We train and test our method on satellite images captured by WorldView II. HRMS images are usually not available in the existing data sets. According to Wald's protocol [33], we downsample the original PAN and MS images into lower resolution and use the original MS image as the HRMS image (ground truth). As the spatial resolution ratio between PAN and LRMS images r is 4, we crop the downsampled PAN image into patches of size $264 \times 264 \times 1$ and the downsampled LRMS images into patches of size $66 \times 66 \times 4$. The original MS images are cropped into patches of size $264 \times 264 \times 4$ as the ground truth. Then, 2052 patch pairs are established as the training data. We set $\lambda = 25$, $\gamma = 20$, $\alpha = 0.5$, and $N = 10$. The model is trained for ten epochs with a batch size of 4. θ_{p} is updated by the AdamOptimizer with a learning rate of 0.002 and exponential decay. The specific training procedure is summarized as Algorithm 1.

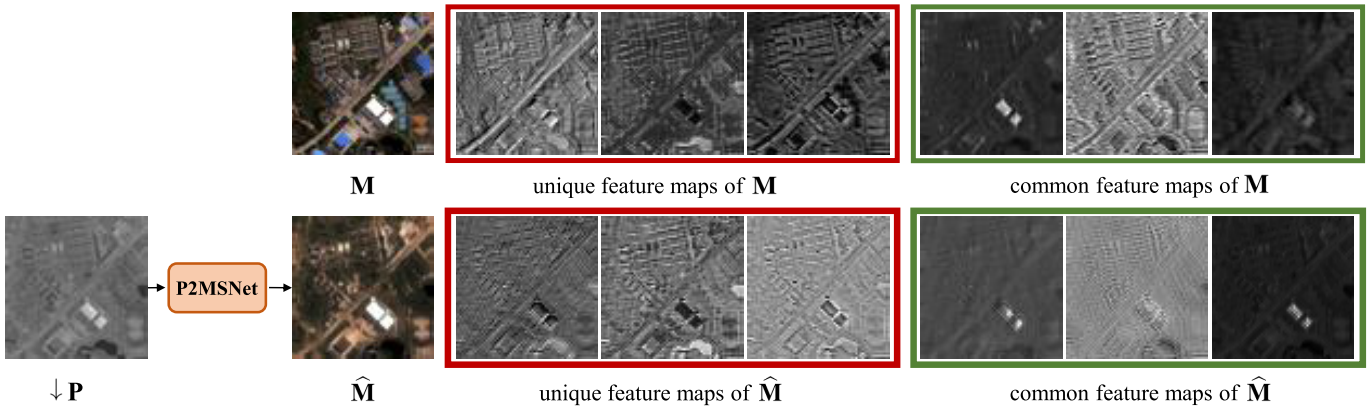

Fig. 6. Illustration of $\mathbf{M}$, $\widehat{\mathbf{M}}$ and some unique/common feature maps extracted from them by the spectral encoder–decoder network.

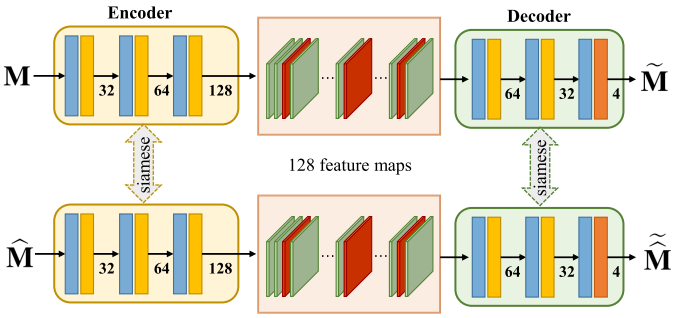

Fig. 7. Network architecture of the spectral encoder–decoder.

B. Results

We compare SDPNet with eight counterparts on the WorldView II data set, including three widely used traditional techniques: PRACS [8], SIRF [34], and LGC [1], which are commonly used algorithms for comparison in pan-sharpening and are usually considered as the representative traditional methods and five deep learning-based methods: PNN [14], PanNet [17], PSGAN [22], TACNN [19], and Pan-GAN [24]. The descriptions of these five deep learning-based methods are given in Section I. For TACNN, as suggested by the authors, we use our training data to perform 5000 epochs of fine-tuning on the trained model provided by the authors. In that case, the network can show better performance.

1) Qualitative Comparison: For qualitative comparison, the visual results on four typical satellite images are presented in Figs. 9–12 (for favorable visualization and perception, the first three bands of MS images are shown). For each group of results, the mean absolute errors (MAEs) between the generated HRMS images and the ground truth are shown on the pseudocolor map in the last row of each figure.

It can be easily observed in these figures that there is severe spectral distortion in the results of SIRF as the deviations in the pixel intensity distributions of the shown three channels are conspicuous. Because the visual result merely shows the corrected result of the first three bands, the spectral distortion cannot be fully reflected. When comparing the MAE, which is computed on all the four bands in the MS images, the results of SIRF suffer from severe spectral distortion. Moreover, as can be seen from the results of PRACS, PanNet, and TACNN, they suffer from severe spatial distortion, represented as blurred details in all the four examples. As for PanNet, there are obvious noises in the results. For PNN and PSGAN, although they can provide clear versions of generated HRMS images visually, there are still several subtle discrepancies in the MAE images compared with our results. The discrepancies between PNN and our proposed SDPNet are perceived in Figs. 9 and 11, and those between PSGAN and our method can also be perceived in these two figures. In Pan-GAN, the spatial and spectral information is preserved by the retention of the gradient and pixel intensity distribution and the adversarial process of the generator and two discriminators. However, the overintroduction of the PAN image gradients and the distinguishing between the pan-sharpened image and the ground truth results in a slight change in the pixel intensity distribution. Due to that LGC aims to utilize the spatial information of the PAN image by designing a constraint on the gradient difference of PAN and HRMS images through a local linear regression model, the spatial structures in the PAN image are well preserved in the results of LGC. However, similar to Pan-GAN, the overreservation of gradient information of the PAN image results in the distortion of the spectral information, which is reflected in the difference between the results of LGC and the ground truth. As can be seen in the highlighted regions in Figs. 9–12, the excessively sharpened or retained gradient information of the PAN images leads to the obvious differences between the results of LGC and the ground truth. By comparison, our SDPNet can achieve an appropriate tradeoff between spectral and spatial preservation.

2) Quantitative Evaluation: For quantitative comparison, five widely used and standard metrics are employed, including four full-reference metrics, i.e., relative dimensionless global error in synthesis (ERGAS) [35], root-mean-squared error (RMSE), relative average spectral error (RASE) [36], and spectral angle mapper (SAM) [37], and a no-reference metric, i.e., spatial correlation coefficient (SCC) [38]. Among these metrics, ERGAS is a global metric that measures the mean shifting and dynamic range change between the result and

Algorithm 1 Training Process of SDPNet Parameter Descriptions: θ_{MS2P} and θ_{P2MS} Are the Parameters in the MS2PNet and P2MSNet to Be Trained Respectively. $\{\theta_{\text{spat\_en}}, \theta_{\text{spat\_de}}\}$ Are the Parameters in the Spatial Encoder and Decoder. $\{\theta_{\text{spec\_en}}, \theta_{\text{spec\_de}}\}$ Are Those in the Spatial Encoder and Decoder. Θ_{p} Are the Parameters in the PNet. The Functions of These Networks Are Represented as f_{MS2P}, f_{P2MS}, f_{spat}, f_{spec}, and f_{p} Correspondingly. The LRMS Image and PAN Image Are Denoted as $\mathbf{M}$ and $\mathbf{P}$, Respectively. The Pan-Sharpened Result Is Denoted as $\mathbf{X}$, and the HRMS Image (Ground Truth) Is Denoted as $\mathbf{G}$

- **Train the MS2PNet and P2MSNet:**
  Initialize θ_{MS2P} and θ_{P2MS}.
  In each training iteration:
  - Sample m PAN patches $\{\mathbf{P}^1, \cdots, \mathbf{P}^m\}$ and m corresponding HRMS patches $\{\mathbf{G}^1, \cdots, \mathbf{G}^m\}$;
  - Obtain the transformed PAN patches with the MS2PNet: $\{\widehat{\mathbf{P}}^1, \cdots, \widehat{\mathbf{P}}^m\} = \{f_{\text{MS2P}}(\mathbf{G}^1), \cdots, f_{\text{MS2P}}(\mathbf{G}^m)\}$;
  - Update θ_{MS2P} with the AdamOptimizer to minimize the loss function defined in Eq. (3) to learn the optimal solution of f_{MS2P};
  - Obtain the transformed HRMS patches with the P2MSNet: $\{\widehat{\mathbf{G}}^1, \cdots, \widehat{\mathbf{G}}^m\} = \{f_{\text{P2MS}}(\mathbf{P}^1), \cdots, f_{\text{P2MS}}(\mathbf{P}^m)\}$;
  - Update θ_{P2MS} with the AdamOptimizer to minimize the loss function defined in Eq. (5) to learn the optimal solution of f_{P2MS};

  θ_{MS2P} and θ_{P2MS} are fixed with no need of training again.
- **Train the spatial encoder and decoder:**
  Initialize $\theta_{\text{spat\_en}}$ and $\theta_{\text{spat\_de}}$.
  In each training iteration:
  - Sample m PAN patches $\{\mathbf{P}^1, \cdots, \mathbf{P}^m\}$ and m corresponding LRMS patches $\{\mathbf{M}^1, \cdots, \mathbf{M}^m\}$;
  - Generate the reconstructed PAN patches: $\{\widetilde{\mathbf{P}}^1, \cdots, \widetilde{\mathbf{P}}^m\} = \{f_{\text{spat}}(\mathbf{P}^1), \cdots, f_{\text{spat}}(\mathbf{P}^m)\}$;
  - Generate the pseudo PAN patches with the pre-trained f_{MS2P}: $\{\widehat{\mathbf{P}}^1, \cdots, \widehat{\mathbf{P}}^m\} = \{f_{\text{MS2P}}(\uparrow \mathbf{M}^1), \cdots, f_{\text{MS2P}}(\uparrow \mathbf{M}^1)\}$;
  - Generate the reconstructed pseudo PAN patches: $\{\widetilde{\widehat{\mathbf{P}}}^1, \cdots, \widetilde{\widehat{\mathbf{P}}}^m\} = \{f_{\text{spat}}(\widehat{\mathbf{P}}^1), \cdots, f_{\text{spat}}(\widehat{\mathbf{P}}^m)\}$;
  - Update $\theta_{\text{spat\_en}}$ and $\theta_{\text{spat\_de}}$ by minimizing the similarity loss between $\{\mathbf{P}^1, \cdots, \mathbf{P}^m\}$ and $\{\widetilde{\mathbf{P}}^1, \cdots, \widetilde{\mathbf{P}}^m\}$ and that between $\{\widehat{\mathbf{P}}^1, \cdots, \widehat{\mathbf{P}}^m\}$ and $\{\widetilde{\widehat{\mathbf{P}}}^1, \cdots, \widetilde{\widehat{\mathbf{P}}}^m\}$;

  $\theta_{\text{spat\_en}}$ and $\theta_{\text{spat\_de}}$ are fixed with no need of training again.
- **Train the spectral encoder and decoder:**
  Initialize $\theta_{\text{spec\_en}}$ and $\theta_{\text{spec\_de}}$.
  In each training iteration:
  - Sample m PAN patches $\{\mathbf{P}^1, \cdots, \mathbf{P}^m\}$ and m corresponding LRMS patches $\{\mathbf{M}^1, \cdots, \mathbf{M}^m\}$;
  - Generate the reconstructed LRMS patches: $\{\widetilde{\mathbf{M}}^1, \cdots, \widetilde{\mathbf{M}}^m\} = \{f_{\text{spec}}(\mathbf{M}^1), \cdots, f_{\text{spec}}(\mathbf{M}^m)\}$;
  - Generate the pseudo LRMS patches with the pre-trained f_{P2MS}: $\{\widehat{\mathbf{M}}^1, \cdots, \widehat{\mathbf{M}}^m\} = \{f_{\text{P2MS}}(\downarrow \mathbf{P}^1), \cdots, f_{\text{P2MS}}(\downarrow \mathbf{P}^m)\}$;
  - Generate the reconstructed pseudo LRMS patches: $\{\widetilde{\widehat{\mathbf{M}}}^1, \cdots, \widetilde{\widehat{\mathbf{M}}}^m\} = \{f_{\text{spec}}(\widehat{\mathbf{M}}^1), \cdots, f_{\text{spec}}(\widehat{\mathbf{M}}^m)\}$;
  - Update $\theta_{\text{spec\_en}}$ and $\theta_{\text{spec\_de}}$ by minimizing the similarity loss between $\{\mathbf{M}^1, \cdots, \mathbf{M}^m\}$ and $\{\widetilde{\mathbf{M}}^1, \cdots, \widetilde{\mathbf{M}}^m\}$ and that between $\{\widehat{\mathbf{M}}^1, \cdots, \widehat{\mathbf{M}}^m\}$ and $\{\widetilde{\widehat{\mathbf{M}}}^1, \cdots, \widetilde{\widehat{\mathbf{M}}}^m\}$;

  $\theta_{\text{spec\_en}}$ and $\theta_{\text{spec\_de}}$ are fixed with no need of training again.
- **Train the PNet:**
  Initialize θ_{p}.
  In each training iteration:
  - Sample m PAN patches $\{\mathbf{P}^1, \cdots, \mathbf{P}^m\}$ and m corresponding LRMS patches $\{\mathbf{M}^1, \cdots, \mathbf{M}^m\}$;
  - Generate the pan-sharpened result patches with the PNet: $\{\mathbf{X}^1, \cdots, \mathbf{X}^m\} = \{f_{\text{p}}(\mathbf{P}^1, \mathbf{M}^1), \cdots, f_{\text{p}}(\mathbf{P}^m, \mathbf{M}^m)\}$;
  - Update θ_{p} by minimizing the loss defined in Eq. (6) by using the pre-trained $\theta_{\text{spat\_en}}$ and $\theta_{\text{spec\_en}}$.

the ground truth. RMSE measures the changes of these two images through the pixel values. In RASE, the difference is reflected through the spectral quality by computing the relative error. SAM measures the spectral similarity between the spectra of the result and that of the ground truth by calculating the angle between them. This angle is computed

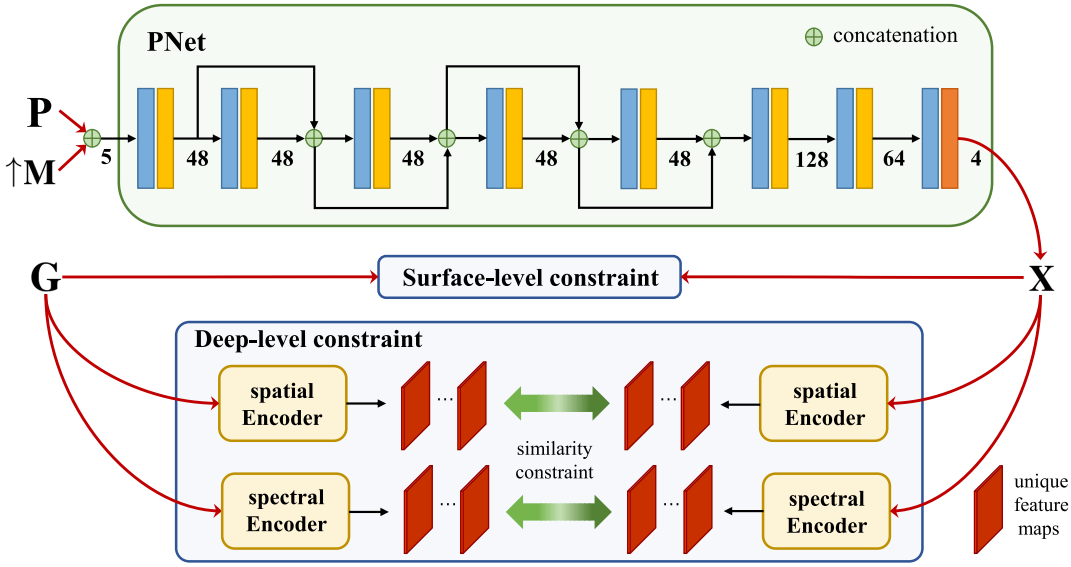

Fig. 8. Network architecture and the constraints of PNet. The spatial encoder and the spectral encoder are those presented in Figs. 5 and 7, respectively. In the testing phase, only the PNet is needed to generate the pan-sharpened HRMS image **X**.

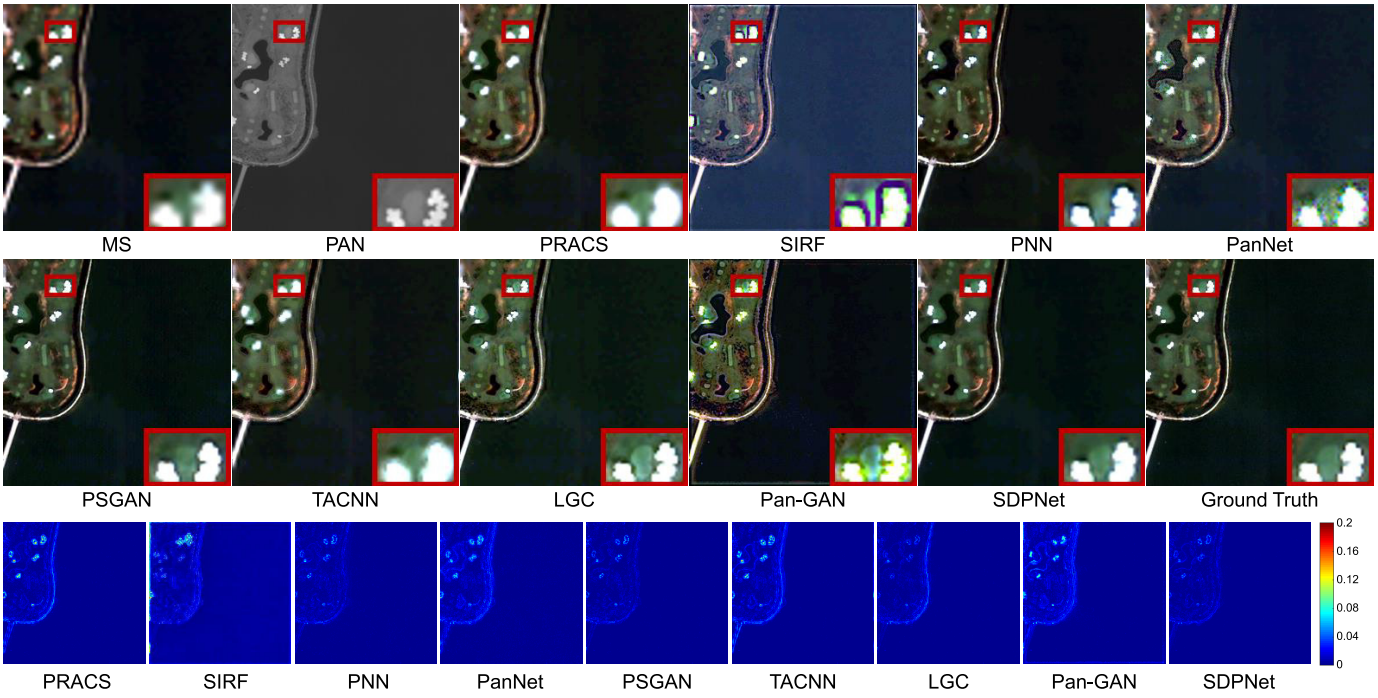

Fig. 9. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set. First row: LRMS and PAN images, pan-sharpened results of PRACS [8], SIRF [34], PNN [14], and PanNet [17]. Second row: results of PSGAN [22], TACNN [19], LGC [1], Pan-GAN [24], SDPNet, and the ground truth. Last row: MAEs between the pan-sharpened results and the ground truth. In the reduce-resolution validation, the pan-sharpened results are of size 264×264.

between the endmember spectrum vector and each pixel vector in n-dimensional space. Smaller angles indicate closer matches to the reference spectrum [39]. SCC is an approach to evaluate the pan-sharpened results without reference. It computes the correlation between the spatial information presented in the PAN image and that of the fused result. A high SCC indicates that much of the spatial detail information of the PAN image is present in the results [40]. Generally speaking, smaller values of ERGAS, RMSE, RASE, and SAM indicate better performance, and larger values of SCC indicate better performance. The average performance and standard deviation across 100 satellite images from WorldView II with different methods are shown in Table I.

As shown in the table, our SDPNet can achieve the best results on the three out of five metrics. As for the remaining metrics, our method can also exhibit the second and third

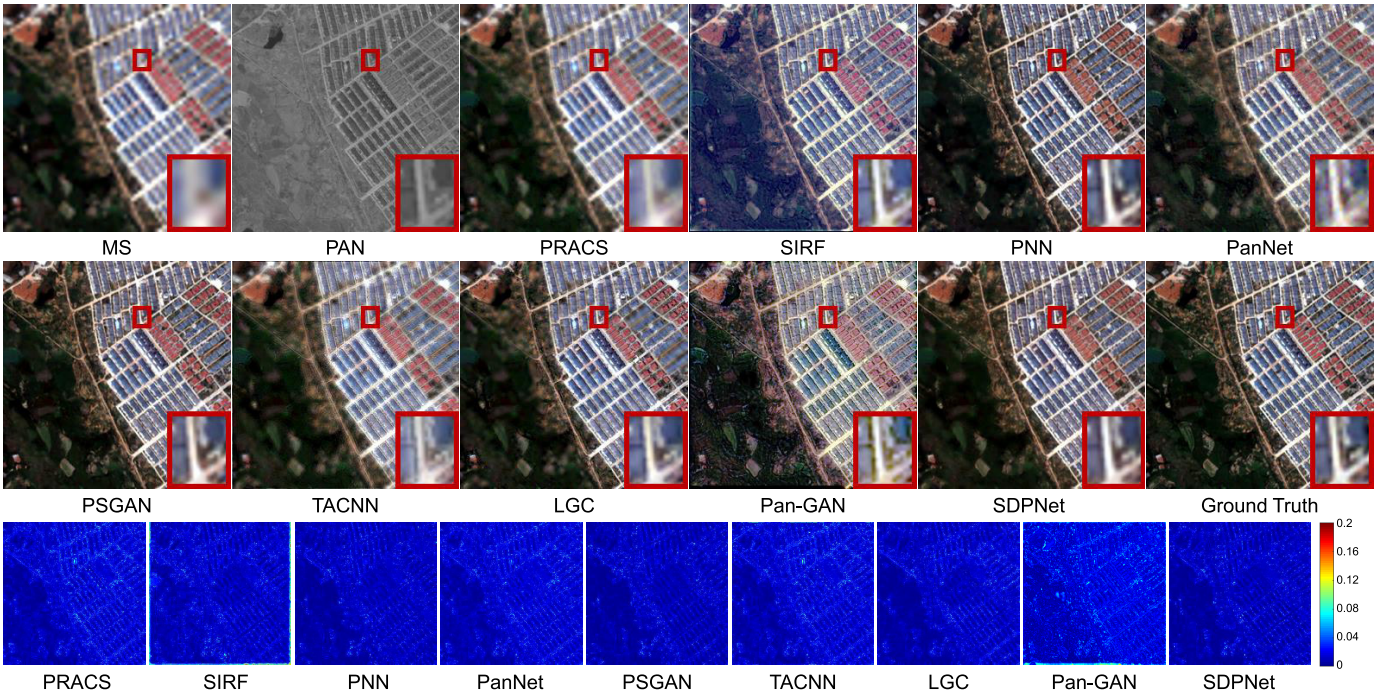

Fig. 10. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set. Images in the last row are the MAEs between the pan-sharpened results and the ground truth.

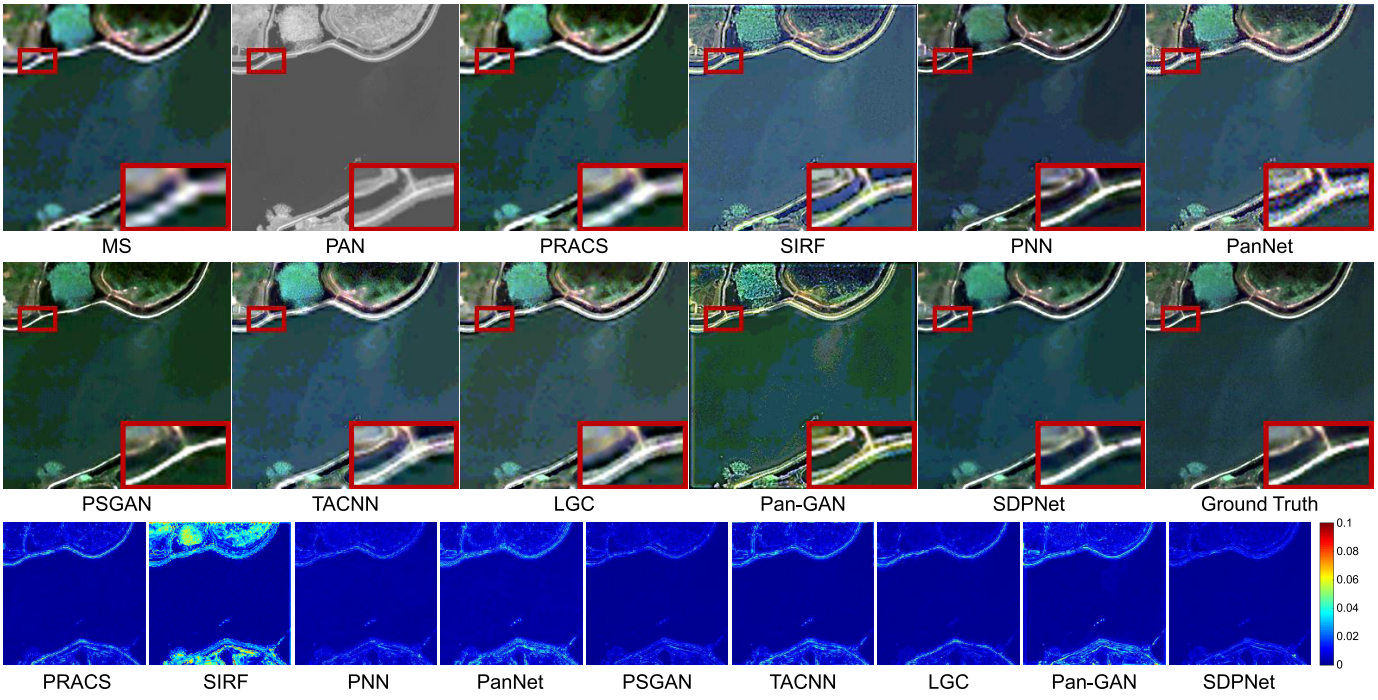

Fig. 11. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set. Images in the last row are the MAEs between the pan-sharpened results and the ground truth.

optimal performances, respectively. More concretely, the best results of our method on ERGAS and RMSE indicate that SDPNet can achieve the least mean shifting and dynamic changes and the least pixel changes between the generated results and the ground truth. The best result of RASE demonstrates that the spectral quality of SDPNet is higher than other counterparts. Besides, the second optimal performance of our method on SAM shows that the spectrum of our results can achieve comparable close matches to the reference one that represents a comparative spectral similarity between our methods and the ground truth. Moreover, SDPNet can achieve the third-best performance on SCC. It shows that SDPNet

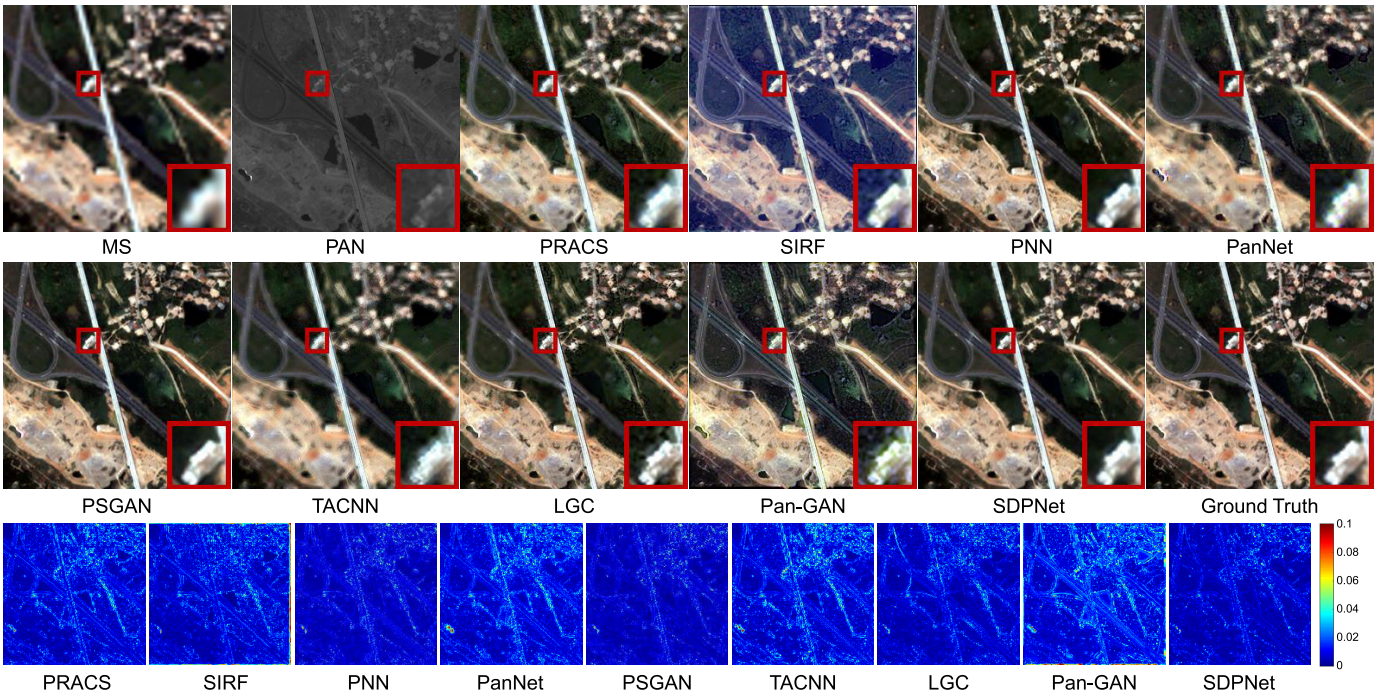

Fig. 12. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set. Images in the last row are the MAEs between the pan-sharpened results and the ground truth.

TABLE I

QUANTITATIVE COMPARISONS ON 100 SATELLITE IMAGES FROM WORLDVIEW II (RED: THE BEST AND BLUE: THE SECOND BEST)

Methods	ERGAS	RMSE	RASE	SAM	SCC
PRACS	1.626 ± 0.385	3.668 ± 1.053	6.628 ± 1.587	2.089 ± 0.587	0.863 ± 0.047
SIRF	3.161 ± 2.555	6.378 ± 2.401	12.564 ± 8.086	3.744 ± 2.966	0.599 ± 0.117
PNN	1.401 ± 0.335	3.086 ± 1.000	5.562 ± 1.443	1.845 ± 0.394	0.843 ± 0.030
PanNet	1.812 ± 0.379	3.976 ± 1.132	7.191 ± 1.614	2.402 ± 0.598	0.733 ± 0.089
PSGAN	**1.280 ± 0.334**	**2.878 ± 1.000**	**5.150 ± 1.456**	**1.583 ± 0.455**	0.812 ± 0.039
TACNN	1.892 ± 0.432	4.224 ± 1.199	7.643 ± 1.851	2.140 ± 0.597	0.837 ± 0.050
LGC	1.312 ± 0.313	2.918 ± 0.965	5.213 ± 1.359	1.715 ± 0.539	**0.904 ± 0.032**
Pan-GAN	2.262 ± 0.485	4.957 ± 1.447	8.938 ± 1.936	2.728 ± 0.782	**0.924 ± 0.027**
SDPNet	**1.256 ± 0.289**	**2.800 ± 0.899**	**5.015 ± 1.266**	**1.603 ± 0.474**	0.868 ± 0.027
desired	0	0	0	0	1

TABLE II

MEAN AND STANDARD DEVIATION OF RUNTIME COMPARISON OF DIFFERENT METHODS ON TEST SATELLITE IMAGES FROM WORLDVIEW II (BOLD: THE BEST AND UNIT: THE SECOND BEST)

	PRACS	SIRF	PNN	PanNet	PSGAN	TACNN	LGC	Pan-GAN	SDPNet
Reduced-resolution	0.12±0.15	27.62±8.21	**0.02±0.04**	0.26±0.07	0.03±0.05	8.17±0.61	11.27±0.87	0.35±0.08	0.10±0.10
Full-resolution	2.10±0.17	835.14±136.38	**0.19±0.03**	0.81±0.13	0.24±0.04	9.13±0.05	548.05±24.47	4.60±0.04	0.51±0.30

can achieve a comparative spatial correlation between the generated results and the PAN images. As for LGC, this method utilizes the spatial information of the PAN image by designing a constraint on the gradient difference of PAN and HRMS images through a local linear regression model. Thus, the spatial structures in the PAN image are well preserved in the results of LGC. Compared with other counterparts, this characteristic is conspicuous and can be seen in all four examples. Thus, LGC can achieve a comparable spatial correlation between the results and the PAN images. However, the spatial distortion is not included in this metric. Thus, through the comprehensive view of the results on all the five metrics, our SDPNet can achieve a satisfactory preservation performance on both the spatial structures and spectral information.

3) Efficiency Comparison: The average runtime and the standard deviation of each method on the 100 satellite images from WorldView II are shown in Table II. The traditional methods are tested on 3.4-GHz Intel Core i5-7500 CPU, and the deep learning-based methods are tested on NVIDIA GeForce GTX Titan X GPU. Compared with traditional methods, deep

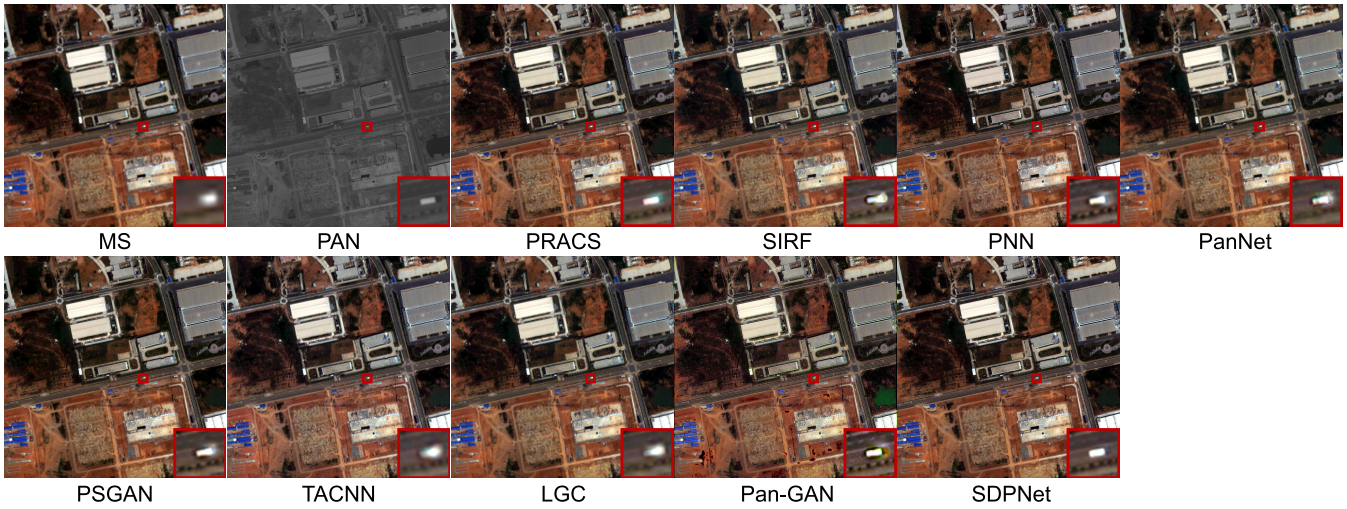

Fig. 13. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set at the original scale. In full-resolution validation, the pan-sharpened results are of size 1056×1056.

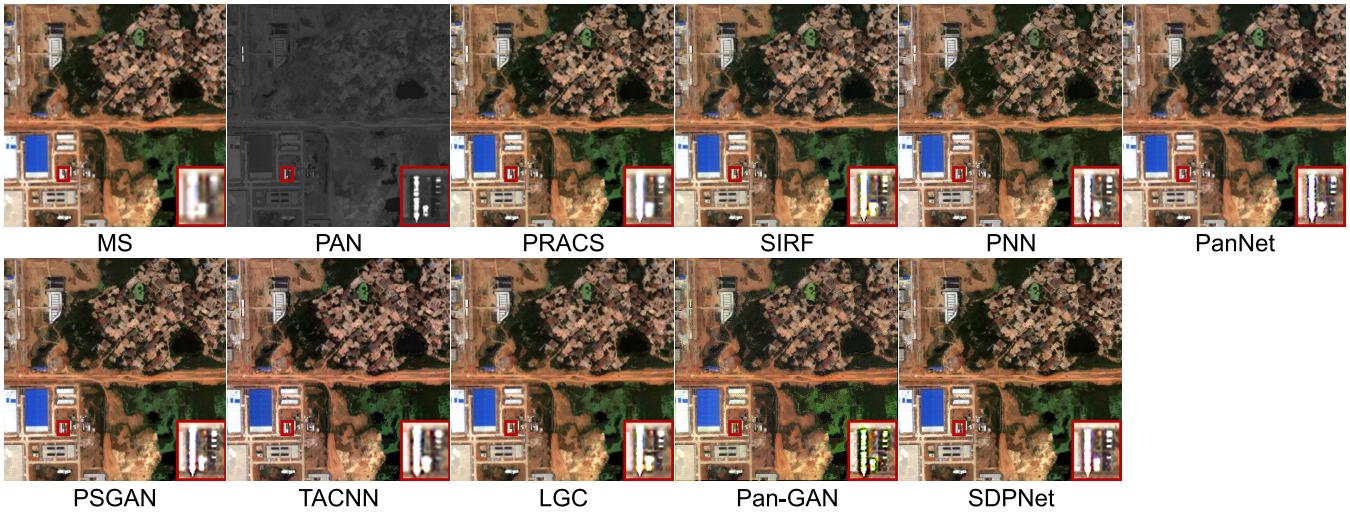

Fig. 14. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set at the original scale.

learning-based methods show superior efficiency. In deep learning-based methods, PNN shows the most superior efficiency because there are only three layers in its architecture and the amount of calculation is relatively small. In the other three methods, the network architectures are all improved. The more computation costs more runtime. By comparison, our method can achieve comparative efficiency under these circumstances.

4) Parameters and Complexity: To analyze the methods more comprehensively, we perform the comparison of the parameter numbers and floating-point operations (FLOPs) [41] in this section. The results are reported in Table III. It can be seen that PNN, TACNN, and Pan-GAN have the minimum number of FLOPs because their network architectures have merely three or five layers. In addition, they only extract a small number of feature maps, resulting in a small number of FLOPs. In PanNet and PSGAN, with the increase in convolutional layers and the increasing complexity of network architecture, both the numbers of parameters and FLOPs increase significantly. In PSGAN, as the size of feature maps in the middle layers is reduced, the increase in parameters does not cause a particularly large increase in the number of FLOPs. However, because there are many basic blocks and convolutional layers in PanNet, the complexity has increased significantly, resulting in a large increase in the number of FLOPs. In our method, the PNet has eight convolutional layers with dense connections and feature maps of more channels, resulting in more parameters and FLOPs. By comparison, our SDPNet improves pan-sharpening performance with more complexity but slightly more parameters.

C. Full-Resolution Validation

In this section, we discuss the full-resolution validation procedure where the PAN and MS images are at the original scale, and thus, the ground-truth images are unavailable. The quantitative comparison results are shown in Figs. 13–15.

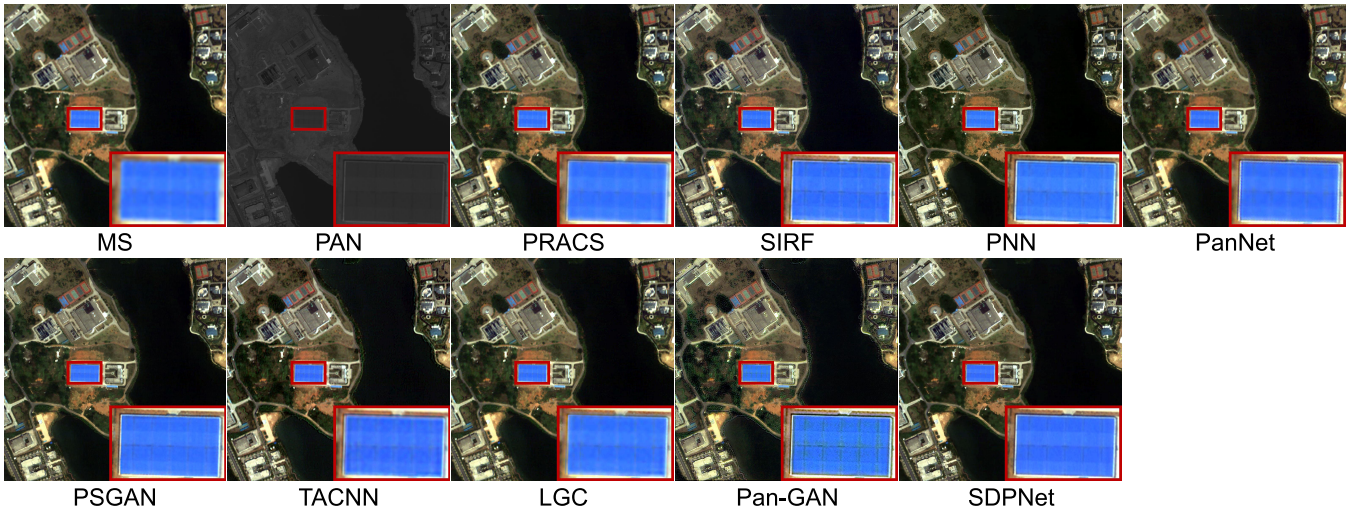

Fig. 15. Qualitative comparison of SDPNet with eight counterparts on a typical satellite image pair from the WorldView II data set at the original scale.

TABLE III
PERFORMANCE COMPARISONS OF DEEP LEARNING-BASED METHODS IN TERMS OF NUMBER OF PARAMETERS (PARAM) AND NUMBER OF FLOPS

	PNN	PanNet	PSGAN	TACNN	Pan-GAN	SDPNet
Param	0.09M	11.19M	1.47M	0.09M	0.09M	0.56M
FLOPs	12.66B	1560.12B	33.88B	13.29B	0.01B	78.44B

For the full-resolution source images, our method shows more obvious advantages. As shown in Fig. 13, the boundaries of the car are blurred in all the competitors, while our result delineates the boundaries. It shows that the spatial structure in the PAN image is well preserved in our results. The qualitative results in Fig. 15 also verify this advantage of the proposed SDPNet. The other advantage can be seen from Fig. 14. While preserving the spatial structure in the PAN image, our result alleviates the spectral distortion, which can be seen from the color differences between the pan-sharpened results and the LRMS images. Given that, in full-resolution validation, the ground truth is unavailable, we use the no-reference metric SCC for quantitative performance evaluation. The results are shown in Table IV. The optimal result of our method on SCC shows that the spatial correlation between our result and the PAN image is high. This quantitative result is consistent with the qualitative results shown in Figs. 13–15 where the spatial structures in our results are similar to those in the PAN images.

The efficiency comparison of different methods on the full-resolution source images is also shown in Table II. When tested on the reduced-resolution source images, the efficiency gaps between different methods are not obvious. However, when the source images are of full resolution, with the increase in spatial scale, the efficiency gaps between the algorithms are further widened. In this instance, our proposed method shows comparative efficiency.

D. Ablation Study

In SDPNet, we constrain the similarity between the generated HRMS images and the ground truth from both surface- and deep-level constraints. The results are generated by introducing the densely connected blocks. In order to verify the effectiveness of the constraints on different levels and the densely connected blocks, we perform the ablation experiments in this section.

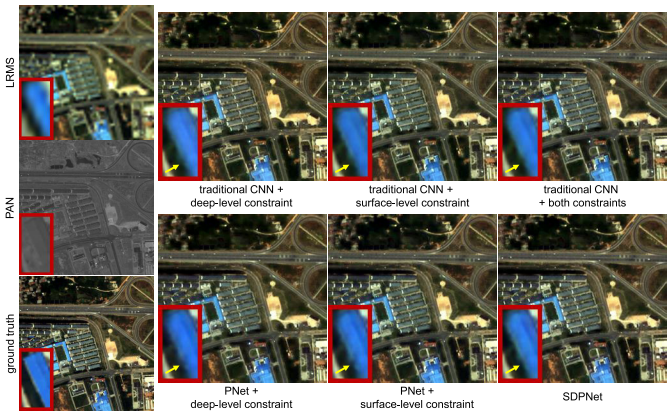

Fig. 16. Qualitative results of applying traditional CNN and PNet with different constraints. The first column: source images (from top to bottom: LRMS, PAN, and ground-truth images). The second to the last columns: pan-sharpened results of applying different network architectures and constraints.

1) Ablation Study of Constraints on Different Levels: To verify the effectiveness of the constraints on surface-level characteristics or deep-level features, only one of them is employed in the ablation experiment. We also compare the results of these three conditions by replacing the PNet with the traditional CNN with the same number of parameters. Both qualitative and quantitative comparisons are performed.

Two representative results of the qualitative comparison of different constraints with PNet can be seen in the second rows of Figs. 16 and 17. As shown in these figures, although the model with a single level constraint, whether surface-level or deep-level, can generate roughly promising results, there is more distortion in their results compared with

TABLE IV

QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON 50 FULL-RESOLUTION SATELLITE IMAGES FROM WORLDVIEW II (RED: THE BEST AND BLUE: THE SECOND BEST)

Methods	PRACS	SIRF	PNN	PanNet	PSGAN	TACNN	LGC	Pan-GAN	SDPNet
SCC	0.865±0.039	0.317±0.132	0.851±0.024	0.709±0.026	0.822±0.030	0.664±0.060	0.819±0.020	**0.866±0.042**	**0.877±0.022**

TABLE V

QUANTITATIVE COMPARISONS OF DIFFERENT CONSTRAINTS ON 100 SATELLITE IMAGES FROM WORLDVIEW II (RED: THE BEST AND BLUE: THE SECOND BEST)

Methods	ERGAS	RMSE	RASE	SAM	SCC
traditional CNN + deep-level	1.465 ± 0.279	3.139 ± 0.905	5.678 ± 1.193	1.812 ± 0.467	0.862 ± 0.025
PNet + deep-level	1.400 ± 0.271	3.121 ± 0.945	5.609 ± 1.284	1.909 ± 0.438	**0.868 ± 0.027**
traditional CNN + surface-level	1.394 ± 0.289	3.095 ± 0.882	5.598 ± 1.152	1.910 ± 0.430	0.696 ± 0.342
PNet + surface-level	1.326 ± 0.302	2.941 ± 0.920	5.293 ± 1.275	1.711 ± 0.467	0.863 ± 0.029
traditional CNN + both	1.283 ± 0.270	2.900 ± 0.874	5.210 ± 1.200	1.698 ± 0.471	0.864 ± 0.028
SDPNet with l_1-loss	**1.277 ± 0.297**	**2.835 ± 0.910**	**5.085 ± 1.284**	**1.632 ± 0.480**	0.867 ± 0.028
SDPNet	**1.256 ± 0.289**	**2.800 ± 0.899**	**5.015 ± 1.266**	**1.603 ± 0.474**	**0.868 ± 0.027**
desired	0	0	0	0	1

Fig. 17. Qualitative results of applying traditional CNN and PNet with different constraints. The first column: source images (from top to bottom: LRMS, PAN, and ground-truth images). The second to the last columns: pan-sharpened results of applying different network architectures and constraints.

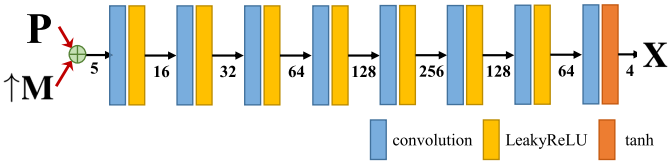

Fig. 18. Network architecture of the traditional CNN. The channel of the feature maps is shown after the LeakyReLU or tanh activation functions.

those of SDPNet. More concretely, it can be reflected in the spatial distortion (blurred boundaries) shown in Fig. 16 and the spectral distortion (color differences) shown in Fig. 17. Moreover, as can be seen from Figs. 16 and 17, the results by merely depending on the deep-level constraint suffer from more severe distortion than merely depending on the surface-level constraint. By comparison, the results of SDPNet exhibit clearer and more similar spatial and spectral information to the ground truth.

To compare the distortion objectively, the quantitative comparison of them is conducted under the abovementioned three conditions. Under different circumstances, the quantitative results on the five metrics mentioned in Section III-B2 are reported in Table V. We see that the single constraint on surface-level characteristics shows better results than the single constraint on deep-level features. By comparison, SDPNet can achieve the best average value on all five metrics. Thus, the combination of both surface- and deep-level constraints can achieve the most superior performance.

2) Ablation Study of PNet: In SDPNet, we also design a network PNet with densely connected blocks to strengthen feature propagation and avoid information distortion. For validation, we replace it with a traditional CNN. The network architecture of PNet is shown in Fig. 18.

The qualitative comparison can be seen in the differences between the results in the first rows and those in the second rows in Figs. 16 and 17. When replacing the traditional CNN with PNet, the differences between the results and the ground truths can be further reduced. In Fig. 16, the improved performance can be seen from the reduced spectral distortion (color differences) of the results in the second row compared with those in the first row correspondingly. In Fig. 17, the results with PNet show less spatial distortion compared with those with CNN, whether in the case of just applying the deep- or surface-level constraint, or both of them.

The quantitative comparison can be seen in Table V. As shown in this table, by replacing the traditional CNN with the PNet under all the three circumstances, i.e., applying deep-, surface-level constraint, and both of them, the results on the five metrics are all improved. To conclude, by combining the advantages of these subparts, i.e., surface- and deep-level constraints and PNet, our method can generate the optimal pan-sharpened results.

3) l_1-Loss Versus l_2-Loss: As described in (3), (5), and (6), we use the combination of the SSIM-based and l_2 losses for surface-level similarity constraint. Given that l_1-loss

is also a commonly used similarity constraint, we perform the experiment where the l_2-loss is replaced by l_1-loss in this section. The quantitative comparison results are shown in Table V. As shown in the results on all the five metrics, the combination of SSIM-based and l_2-losses shows better performance by achieving a higher similarity between the generated and reference images. Although l_2-loss suffers from relatively blurred results by averaging all plausible outputs, the combination of it and SSIM-based loss can alleviate this shortcoming as SSIM focuses on the structure information. By comparison, l_1-loss tends to produce a sparse solution and has a certain tolerance for implausible outputs. Thus, the combination of l_2-loss rather than l_1-loss with SSIM is more suitable for similarity constraint.

IV. Conclusion

In this article, we have proposed a new deep network based on surface- and deep-level constraints, termed SDPNet, to address the pan-sharpening problem. For further spatial and spectral information preservations, we first design two encoder–decoder networks to extract deep-level features from two types of source images, in addition to surface-level characteristics to enhance the information representation. The feature maps that characterize the unique information in the PAN or LRMS image can be regarded as spatial or spectral-related feature maps. The pan-sharpened result is supposed to exhibit a similar spatial- or spectral-related feature maps with the ground truth. Thus, the similarity between them can be further increased with less information distortion. Compared with state-of-the-art methods with both reduced-resolution and full-resolution validations, our method can produce pan-sharpened results with less spatial and spectral distortions.

References

[1] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10265–10274.

[2] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *Proc. GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion Over Urban Areas*, 2003, pp. 90–94.

[3] N. H. Kaplan and I. Erer, "Bilateral pyramid based pansharpening of multispectral satellite images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 2376–2379.

[4] V. P. Shah, N. H. Younan, and R. King, "Pan-sharpening via the contourlet transform," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2007, pp. 310–313.

[5] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[6] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[7] 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.

[8] 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.

[9] X. Kang, S. Li, and J. A. Benediktsson, "Pansharpening with matting model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5088–5099, Aug. 2014.

[10] M. Ghahremani and H. Ghassemian, "A compressed-sensing-based pan-sharpening method for spectral distortion reduction," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2194–2206, Apr. 2016.

[11] 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.

[12] C. Kwan, J. Choi, S. Chan, J. Zhou, and B. Budavari, "A super-resolution and fusion approach to enhancing hyperspectral images," *Remote Sens.*, vol. 10, no. 9, p. 1416, Sep. 2018.

[13] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 28, 2020, doi: 10.1109/TPAMI.2020.3012548.

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

[15] 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.

[16] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sens. Imag.*, vol. 17, no. 1, p. 10, Dec. 2016.

[17] 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. 1753–1761.

[18] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[19] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[20] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[21] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[22] X. Liu, Y. Wang, and Q. Liu, "Psgan: A generative adversarial network for remote sensing image pan-sharpening," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 873–877.

[23] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, and Y. Zhang, "Residual encoder–decoder conditional generative adversarial network for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1573–1577, Sep. 2020.

[24] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020.

[25] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.

[26] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12797–12804.

[27] M. Selva, L. Santurri, and S. Baronti, "On the use of the expanded image in quality assessment of pansharpened images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 320–324, Mar. 2018.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[29] M. Decombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, and F. Capman, "A new object based quality metric based on SIFT and SSIM," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1493–1496.

[30] J.-F. Pambrun and R. Noumeir, "Limitations of the SSIM quality metric in the context of diagnostic imaging," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2960–2963.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[32] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 902–911.

[33] 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.

[34] 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.
[35] 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.
[36] M. Choi, "A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1672–1682, Jun. 2006.
[37] 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.
[38] 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, Jan. 1998.
[39] H. Z. M. Shafri, A. Suhaili, and S. Mansor, "The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis," *J. Comput. Sci.*, vol. 3, no. 6, pp. 419–423, Jun. 2007.
[40] 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.
[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

Han Xu received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018, where she is pursuing the Ph.D. degree with the Multi-spectral Vision Processing Laboratory, Electronic Information School.

She has first-authored several refereed journal articles and conference papers, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE/the IEEE TRANSACTIONS ON IMAGE PROCESSING, AAAI Conference on Artificial Intelligence, the International Joint Conferences on Artificial Intelligence, and so on. Her research interests include computer vision and remote sensing.

Jiayi Ma received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

He is a Professor with Electronic Information School, Wuhan University, Wuhan. He has authored or coauthored more than 140 refereed journal articles and conference papers, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE/the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *International Journal of Computer Vision*, the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, European Conference on Computer Vision, and so on. His research interests include computer vision, machine learning, and pattern recognition.

Dr. Ma has been identified in the 2019 highly cited researchers' list from the Web of Science Group. He is an Area Editor of *Information Fusion*, an Associate Editor of *Neurocomputing*, and a Guest Editor of *Remote Sensing*.

Zhenfeng Shao (Member, IEEE) received the Ph.D. degree in aerial photogrammetry from Wuhan University, Wuhan, China, in 2004.

He is a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include remote sensing and data mining.

Hao Zhang received the B.E. degree from the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China, in 2019. He is pursuing the master's degree with Electronic Information School, Wuhan University, Wuhan.

His research interests include computer vision, machine learning, and pattern recognition.

Junjun Jiang (Member, IEEE) received the B.S. degree in information and computing science from the School of Mathematical Sciences, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree in communication and information system from the School of Computer, Wuhan University, Wuhan, China, in 2014.

He is a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include applications of image processing and pattern recognition in video surveillance, image super-resolution, image interpolation, and face recognition.

Xiaojie Guo (Senior Member, IEEE) is a tenured Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China.

Dr. Guo was a recipient of the Piero Zamperoni Best Student Paper Award in the International Conference on Pattern Recognition in 2010 and the IEEE ICME Best Student Paper Runner-Up Award in 2018.