LDP-Net: An Unsupervised Pansharpening Network Based on Learnable Degradation Processes

Jiahui Ni, Zhimin Shao, Zhongzhou Zhang, Mingzheng Hou, Jiliu Zhou, *Senior Member, IEEE*,
Leyuan Fang, *Senior Member, IEEE*, and Yi Zhang, *Senior Member, IEEE*

***Abstract*—Pansharpening in remote sensing image aims at acquiring a high-resolution multispectral (HRMS) image directly by fusing a low-resolution multispectral (LRMS) image with a panchromatic (PAN) image. The main concern is how to effectively combine the rich spectral information of LRMS image with the abundant spatial information of PAN image. Recently, many methods based on deep learning have been proposed for the pansharpening task. However, these methods usually have two main drawbacks: 1) requiring HRMS for supervised learning; and 2) simply ignoring the latent relation between the MS and PAN image and fusing them directly. To solve these problems, we propose a novel unsupervised network based on learnable degradation processes, dubbed as LDP-Net. A reblurring block and a graying block are designed to learn the corresponding degradation processes, respectively. In addition, a novel hybrid loss function is proposed to constrain both spatial and spectral consistency between the pansharpened image and the PAN and LRMS images at different resolutions. Experiments on GaoFen-2, Worldview-2, and Worldview-3 images demonstrate that our proposed LDP-Net can fuse PAN and LRMS images effectively without the help of HRMS samples, achieving promising performance in terms of both qualitative visual effects and quantitative metrics.**

***Index Terms*—Image fusion, pansharpening, remote sensing, unsupervised learning.**

I. Introduction

NOWADAYS, numerous remote sensing images are obtained to monitor the conditions of agriculture, forestry, ocean, land, environmental protection, and meteorology [1]. Usually, most earth observation satellites can provide two kinds of images, namely, panchromatic (PAN) images with a high spatial resolution band and multispectral (MS) images with higher spectral resolution but lower spatial resolution, which are limited to the image signal-to-noise ratio (SNR) and data storage and transmission. Naturally, the technique for PAN and MS image fusion has been proposed and developed. This technology, which is known as pansharpening, integrates the complementary advantages of spatial and spectral information respectively from PAN and MS images to obtain high spatial resolution MS images. Fused images with both high spectral and spatial resolution can achieve better results in subsequent tasks, such as image classification and object detection [2].

In early research, many traditional methods were proposed to develop pansharpening algorithms, and most of them can be generally summarized into three categories.

1) 1) Methods based on component substitution (CS) [3] attempt to transform MS images and PAN images into a new space in which the structural component of MS images can be substituted by PAN images to achieve spatial information injection. Representative attempts include principal component analysis (PCA) [4], intensity-hue-saturation (IHS) [5], and Gram–Schmidt adaptive (GSA) transform [6].
2) Multiresolution-analysis-based methods utilize the high frequencies of PAN images to restore the spatial details in MS images. To extract this high-frequency information in PAN images, various transform algorithms are applied, such as Laplacian pyramid transform [7], discrete wavelet transform (DWT) [8], and support value transform [9].
3) Model-based methods [10] treat pansharpening as an inverse process of the degradation in which the ideal high-resolution multispectral (HRMS) image degenerates to a PAN image and low-resolution multispectral (LRMS) image. One typical example is the band-dependent spatial detail (BDSD) method [11].

However, these methods exhibit a certain degree of spectral distortions owing to some prior assumptions, which are hard to be generalized to different situations [12].

In the past decade, deep learning approaches, especially convolutional neural networks (CNNs), have achieved excellent performance in various fields, including computer vision and image processing tasks [13]. Some pioneering methods have applied CNNs to the pansharpening task. Typical examples include PNN [14], PanNet [15], PSGAN [16], RED-cGan [17], and TFNet [18]. These supervised learning methods use an end-to-end network to learn the pansharpening process and achieve desirable performance with high spatial resolution and few spectral distortions. However, two vital problems still exist in most CNN-based methods. The first issue is that most networks are based on supervised learning and the training data

Manuscript received 22 April 2022; revised 8 June 2022; accepted 26 June 2022. Date of publication 4 July 2022; date of current version 18 July 2022. This work was supported in part by Sichuan Science and Technology Program under Grant 2021JDJQ0024, in part by Sichuan University "From 0 to 1" Innovative Research Program under Grant 2022SCUH0016, in part by the National Natural Science Fund of China under Grant 61922029, and in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2019RS2016. *(Corresponding author: Yi Zhang.)*

Jiahui Ni is with the School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China, and also with the College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China (e-mail: jiahui_ni@stu.scu.edu.cn).

Zhimin Shao, Zhongzhou Zhang, Mingzheng Hou, and Jiliu Zhou are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: shaozm_3@foxmail.com; zz_zhang@stu.scu.edu.cn; houmingzheng@scu.edu.cn; zhoujl@scu.edu.cn).

Leyuan Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: fangleyuan@gmail.com).

Yi Zhang is with the School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China, and also with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: yzhang@scu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3188181

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

are generated following Wald's protocol [19]. These models perform spatial downsampling and blurring operations on the MS images to obtain the LRMS images and treat the original MS images as ground truth. These operations may not be consistent with the degradation processes in the real situation. The other issue is that these schemes do not effectively utilize the rich spatial information of PAN images [20] and ignore the relation between MS images and PAN images.

To address these problems, we propose a novel unsupervised network for pansharpening based on a two-stream CNN-based architecture with two learnable degradation processes, dubbed as LDP-Net. Pansharpening can be regarded as a superresolution or deblurring problem [21] with additional PAN images and aims to restore the spatial details from PAN images and simultaneously maintain the spectral information of LRMS images. Owing to the lack of ground truth, the inverse process of pansharpening can be divided into two degradation processes: one process uses a spectral response function to transform the HRMS image into a single grayed image similar to the PAN image, and the other process models a spatial blurring operation from the HRMS image into an upsampled LRMS image with a blurring kernel. In the proposed LDP-Net, we adopt two CNN modules to learn two degradation processes. Moreover, according to the relation between MS and PAN images, we propose a new loss function to effectively constrain both spatial and spectral information. Furthermore, a KL divergence loss function is proposed to maintain the spectral distribution of the difference between the MS and PAN images at two resolutions, which has never been explored. As a result, our proposed model achieves desirable performance in which the predicted HRMS image can preserve the high spatial resolution of the PAN image and rich spectral information of the LRMS image under unsupervised conditions. The main contributions of this article are summarized as follows.

1) An unsupervised pansharpening model is proposed based on a two-stream end-to-end network, which is trained without relying on supervised labels. The hyperparameters of the model can be easily tuned in training phase.
2) Different from other models with specified degradation operators, our proposed model learns the degradation processes in a data-driven manner.
3) A novel hybrid loss function, which consists of three parts, is proposed. The first two parts maintain the spatial and spectral consistency between the inputs and the predicted HRMS image in two different resolutions. The other part constrains the difference between the MS and PAN images at different resolutions to have similar distributions.
4) Extensive experiments on different remote sensing datasets demonstrate the effectiveness and robustness of our method over several state-of-the-art methods in both qualitative and quantitative aspects.

The rest of this article is organized as follows. In Section II, we review related works on pansharpening. Section III introduces the framework of the proposed unsupervised model and the loss function for training without labels. In Section IV, extensive experiments were conducted to illustrate our pansharpening method compared with several representative traditional, supervised, and unsupervised learning based approaches. Finally, Section V concludes this article.

II. Related Works

Numerous pansharpening methods have emerged in recent decades, and this section briefly reviews these methods, including classic approaches, supervised learning based approaches and unsupervised learning based approaches.

A. Classic Methods

Traditional pansharpening methods can be roughly classified into three categories. First, early pansharpening studies focused on CS. Some components of the upsampled LRMS images are substituted by corresponding components of PAN images in a specific transform domain. The spectral information and spatial information are separated using a simple and fast transformation, such as IHS [5], principal components transform [22], and GSA transform [6]. Moreover, Dou *et al.* [23] proposed a general framework to implement these CS-based methods systematically. These methods can effectively achieve high spatial resolution but may cause spectral distortions in the pansharpened results. The second category is multiresolution-analysis-based methods, which apply multiscale decomposition techniques to inject high-frequency information of the PAN image into the upsampled LRMS image. High-frequency spatial information is usually extracted by several transform algorithms, such as wavelet transform [24], Laplacian pyramid transform [7], curvelet transform [25], and contourlet transform [4]. Although these methods can achieve improved performance in spectral fidelity, they may also cause aliasing distortion and blurring effects in spatial details. The third type is model-based methods. For instance, Garzelli *et al.* [11] presented two linear injection models, including the single spatial detail (SSD) model and the BDSD model and optimized the models by minimizing the squared error between the original MS image and pansharpened results. Another pansharpening model proposed by Wright achieved fast image fusion with a Markov random field [26]. In addition, Guo *et al.* [27] adopted an online coupled dictionary learning approach to model the relation between LRMS and PAN images to reduce the spectral distortion and restore the spatial details. Recently, Guo *et al.* [28] developed a new posterior probability model based on the Bayesian theory to achieve better spectral and spatial fusion.

B. Supervised Learning Based Approaches

These deep learning methods specifically design a CNN-based network driven by large quantities of paired training data and achieve better performance than traditional methods. Motivated by the superresolution convolutional neural network (SRCNN) model [29], Giuseppe *et al.* [14] first proposed a three-layer CNN-based network named PNN according to the characteristics of remote sensing images. Later, Yang *et al.* [15] directly added the upsampled LRMS image to the output of the network to maintain spectral consistency and treated the edges of PAN and LRMS images as the inputs of the network to restore the spatial details. However, introducing only high-frequency information and superimposing the upsampled LRMS image on the results can cause a blurring effect and lead the training difficult to converge. Scarpa *et al.* [30] adopted a target-adaptive usage modality to ensure that a lightweight network can be applied to different remote sensing sensors. In the deep residual pansharpening neural network (DRPNN) model [31], the concept of residual learning is introduced to form a very deep convolutional neural network, which can further improve the pansharpening performance. He *et al.* [32] introduced a new detail injection strategy into the CNN-based pansharpening methods. Subsequently, Deng *et al.* [33] further exploited a new detail injection-based network aided by the difference between the PAN image and the upsampled LRMS image. Recently, Liu *et al.* [18] also incorporated residual learning into a two-stream CNN architecture to fuse the features extracted from both MS and PAN images. Zhang *et al.* [34] designed a triple-double network with a level-domain-based loss function to fully exploit the spatial details of the PAN image. Jin *et al.* [35] utilized Laplacian pyramid network to recover the crucial spatial information at multiscales. Moreover, several generative adversarial network (GAN)-based methods have been proposed to utilize a discriminator to distinguish the generated images from the ground-truth images. In PSGAN [16], the authors first attempted to produce high-quality pansharpened images with GANs and design a two-stream fusion architecture as the generator and a fully convolutional network as the discriminator. In RED-cGAN [17], a residual encoder–decoder conditional GAN was proposed to produce more details with sharpened images. However, as we mentioned above, these methods require HRMS images for supervised learning and still suffer from spectral distortions or blurring effects.

C. Unsupervised Learning Based Approaches

To address the unreality of simulated data and bridge the gap between classic and supervised learning based approaches, some unsupervised learning based approaches have been developed. Ma *et al.* [20] achieved unsupervised pansharpening using one generator and two discriminators that were designed to distinguish the spatial and spectral characteristics between generated and real images, respectively. Then, Zhou *et al.* [36] combined a generative multiadversarial network and nonreference loss function to improve the performance of unsupervised pansharpening. Motivated by some priors about downsampling and blurring, several methods have been developed for unsupervised pansharpening. For instance, a deep learning prior based on spatial downsampling with blurring has been applied for image fusion to obtain the loss function in [37]. The authors embedded the semantic features extracted from the guidance PAN image by an encoder–decoder network into another deep decoder to generate an output image. Similarly, Luo *et al.* [38] designed an iterative network architecture with a PAN-guided strategy and a set of skip connections to continuously extract and fuse the features from the input and then used a fixed unidimensional Gaussian kernel to obtain a blurred version from the fused HRMS image. However, these prior-based methods are limited to handcrafted training data and cannot be effectively applied to real scenes.

In this article, we propose an unsupervised learning model based on a two-stream CNN network incorporated with two learnable degradation modules that can be adaptive to complex simulated and real situations. Moreover, we specifically design a hybrid spectral loss to effectively maintain spectral consistency between the output and input LRMS images.

III. Method

A. Problem Formulation and Framework

Unsupervised pansharpening aims to obtain the pansharpened HRMS image by fusing the LRMS image and the HR PAN image without the ground-truth. We denote the LRMS image by $m \in R^{w\times h\times C}$, the corresponding HR PAN image by $P \in R^{W\times H}$, the pansharpened HRMS image by $\widehat{M} \in R^{W\times H\times C}$, and the ground-truth HRMS image by $M \in R^{W\times H\times C}$. W and H represent the width and height of high-resolution images, respectively, while w and h represent the width and height of low-resolution images, respectively. C is the number of spectral bands of the multispectral image and usually, $C = 4$. The scale factor for spatial resolution ratio is defined as $r = W/w = H/h$ and usually $r = 4$.

Our proposed LDP-Net is based on a two-stream encoder–decoder fusion network. As shown in Fig. 1, the network mainly consists of several different modules, including feature extraction block (FEB), dense encoder–decoder block (DEDB), reconstruction block (REC), graying block (GB), and reblurring block (RB). First, we interpolate the LRMS image m to the upsampled LRMS image $\uparrow m \in R^{W\times H\times C}$ with same resolution as that of the PAN image. As shown in Fig. 1, to unify the dimensions of both HR PAN image P and the predicted HRMS image $\widehat{M}$ as the input of RB, we copy the single-band PAN image C times to form a C-band tensor $\widetilde{P} \in R^{W\times H\times C}$. Then, $\uparrow m$ and $\widetilde{P}$ are fed into FEB to obtain the shallow spectral and spatial features F_m and F_p, respectively. Then, we use DEDB [39], which has a strong inference ability to further extract and fuse the deep features. Finally, the predicted HRMS image $\widehat{M}$ is reconstructed from the concatenation of the deep features and shallow features via two residual connections [13]. The fusion process takes the following general form:

$$\widehat{M} = f\left(\uparrow m, \widetilde{P}; \Theta\right) \tag{1}$$

where $f(\cdot)$ is the two-stream encoder–decoder fusion model, which takes $\uparrow m$ and $\widetilde{P}$ as the inputs and generates the desired HRMS image $\widehat{M}$, while Θ is the collection of parameters for this model.

Since we do not have the HRMS image as labels, to achieve unsupervised learning, two degradation processes, namely, the degradation between the ideal HRMS image M and the HR PAN image P and the degradation between the ideal HRMS image M and the upsampled LRMS image $\uparrow m$, are formulated to add extra constraints on the training procedure of f as follows [40]:

$$P = \sum_{i=1}^{C} \alpha_i M_i \tag{2}$$

and

$$\uparrow m = k * M \tag{3}$$

where M_i denotes the ith band of the ideal HRMS image, α_i is the corresponding weighting coefficient, and k represents the spatial blur kernel. (2) can be regarded as the degradation process of graying an MS image, which is similar to graying a RGB image, while (3) can be regarded as the blurring process. Inspired by the forms of (2) and (3), a channel attention module [41] is adopted to simulate the graying degradation as GB and a convolution module is employed to simulate the blurring degradation as RB. The parameters of both modules can be learned from training data. Consequently, our model can be optimized by minimizing the loss between the inputs and two degraded versions of the output HRMS image $\widehat{M}$. In addition, we apply the two degradation blocks for the two inputs to obtain their corresponding degraded versions at low resolution. It is worth mentioning that the different GBs and RBs used in our proposed LDP-Net share the same parameters, respectively.

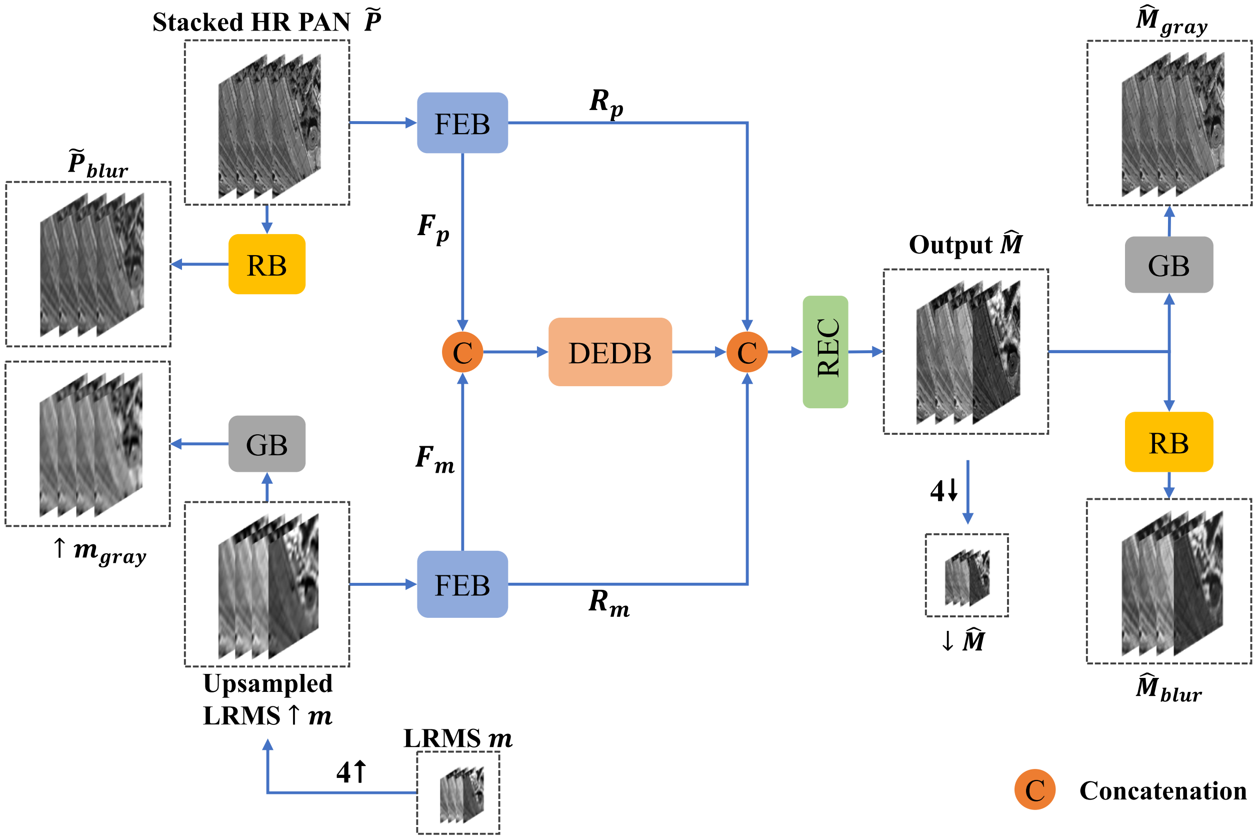


Fig. 1. Overview of the proposed LDP-Net for pansharpening. FEB denotes the feature extraction block. DEDB denotes the dense encoder–decoder block. RB and GB represent the reblurring block and graying block, respectively. REC stands for the reconstruction block. 4 ↑ and 4 ↓ stand for 4 times upsampling and downsampling, respectively. F and R denote the shallow features and residual connection, respectively.

B. Loss Function

Given the upsampled LRMS image $\uparrow m$ and the stacked HR PAN image $\widetilde{P}$ as the inputs, our network produces the desired HRMS image $\widehat{M}$ and four degraded images using the learned degradation operations, which are respectively defined as follows:

$$\widehat{M}_{gray} = G\left(\widehat{M}\right) \tag{4}$$

$$\widehat{M}_{blur} = B\left(\widehat{M}\right) \tag{5}$$

$$\uparrow m_{gray} = G\left(\uparrow m\right) \tag{6}$$

$$\widetilde{P}_{blur} = B\left(\widetilde{P}\right) \tag{7}$$

where $\widehat{M}_{gray} \in R^{W\times H\times C}$ is the grayed version of $\widehat{M}$, $\widehat{M}_{blur} \in R^{W\times H\times C}$ is the blurred version of $\widehat{M}$, $\uparrow m_{gray} \in R^{W\times H\times C}$ denotes the grayed version of $\uparrow m$, and $\widetilde{P}_{blur} \in R^{W\times H\times C}$ denotes the blurred version of $\widetilde{P}$. $G(\cdot)$ and $B(\cdot)$ represent the corresponding degradation functions of GB and RB, respectively. Then, our model utilizes these degraded versions to calculate the loss without the ground-truth. The proposed loss function contains three parts: spatial loss, spectral loss, and spectral KL divergence loss.

1) Spatial Loss: The degradation relationship between the MS image and PAN image can be used to restore the high-resolution spatial information of the output HRMS image. Thus, the spatial loss of our method, which can be divided into spatial constraints at both low and high resolutions, is defined as

$$L_{spatial} = \left\|\widetilde{P}_{blur} - \uparrow m_{gray}\right\|_2^2 + \delta * \left\|\widetilde{P} - \widehat{M}_{gray}\right\|_2^2 \tag{8}$$

where $\|\cdot\|_2$ denotes the L2 norm and δ represents a regularization parameter to balance the two terms. The first term represents the spatial constraint at low resolution, and the second term represents the spatial constraint at high resolution after upsampling. The proposed spatial loss devotes to ensuring the consistency of spatial information extracted by two degradation modules at different resolutions.

2) Spectral Loss: Another degradation between the HRMS image and the upsampled LRMS image can be regarded as the blurring operation, which can be used to maintain the spectral consistency between the output HRMS image and the input upsampled LRMS image at different resolutions. Then, similar to (8), the spectral loss is defined as

$$L_{spectral} = \left\|\uparrow m - \widehat{M}_{blur}\right\|_2^2 + \gamma * \left\|m - \downarrow \widehat{M}\right\|_2^2 \tag{9}$$

where γ denotes a regularization parameter to balance the two terms. To get the downsampled output $\downarrow \widehat{M}$, we smooth $\widehat{M}$ with 3×3 mean filter before downsampling to avoid aliasing effect.

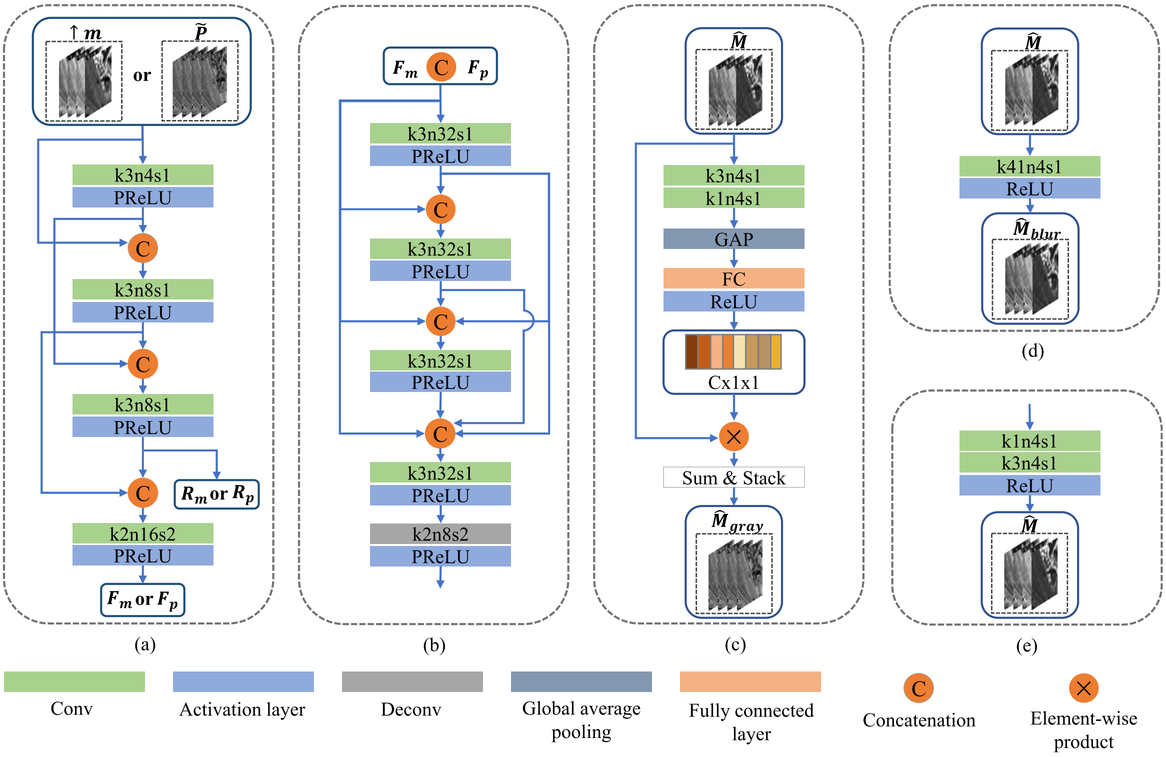


Fig. 2. Structure of (a) FEB, (b) DEDB, (c) GB, (d) RB, and (e) REC, where k3n128s1 denotes a convolution layer with a 3 × 3 kernel size, 128 channels, and stride 1.

3) Spectral KL Divergence Loss: On the other hand, we consider the inverse process of graying degradation and note that the spectral information of MS images in different spectral bands should follow a specific pattern. The difference between the MS image and PAN image at different resolutions should have similar distributions. Based on this consideration, we use the softmax function to transform the residual terms into a form of probability distribution. Then, the spectral Kullback–Leibler (KL) divergence loss is added to constrain the distribution of the residual terms at different resolutions, which is formulated as follows:

$$L_{KL} = KL(p(x_{low})\,\|q(x)) \ , \tag{10}$$

where $p(x_{low}) = softmax(\uparrow m - \uparrow m_{gray})$ and $q(x) = softmax(\widehat{M} - \widetilde{P})$. $x_{low} = \uparrow m - \uparrow m_{gray}$ denotes the residual features between the MS image and the PAN image at low resolution and $x = \widehat{M} - \widetilde{P}$ stands for the residual features between the MS image and the PAN image at high resolution. We reshape the residual terms into a one-dimensional vector and apply the softmax function to rescale the elements. Then KL divergence is applied to impose both terms have similar distributions. The spectral KL divergence loss can effectively reduce the spectral artifacts in the fused results, which will be demonstrated in the experimental section.

In summary, we utilize spatial loss and spectral loss to simultaneously restore the spatial details and preserve the spectral information from the inputs. Moreover, an additional spectral KL divergence loss is proposed to further adjust the spectral qualities. Finally, our proposed unsupervised model is trained by minimizing the following loss function:

$$L = \alpha L_{spatial} + \beta L_{spetral} + \mu L_{KL}. \tag{11}$$

where α, β, and μ are the weights that are empirically set in our experiments. It can be seen that the proposed loss function can be used to train the proposed LDP-Net without the HRMS image (ground-truth) via two degradation processes that can learn the latent characteristics of the output HRMS image.

C. Network Architecture

As mentioned in Section III-A, there are several CNN-based blocks that are designed to implement our proposed network framework, including FEB, DEDB, GB, and RB. Specifically, FEB is used to extract the shallow features from the upsampled LRMS image and HR PAN image to contribute to the subsequent fusion step. Thus, given $\uparrow m$ or $\widetilde{P}$ as the inputs, the corresponding shallow features F_m or F_p can be obtained as

$$F_m = f_{FEB}(\uparrow m) \tag{12}$$

and

$$F_p = f_{FEB}(\widetilde{P}) \tag{13}$$

where f_{FEB} represents the operation of FEB. It must mention that both blocks have the same structure but different parameters that extract different features from the MS and PAN images, respectively. As shown in Fig. 2(a), three convolutional layers with several adjacent residual connections are adopted to extract features from different depths and one downsampling convolutional layer is used to reduce the size of features. As shown in Fig. 1, R_p and R_m denote the output of the residual connection

TABLE I
PARTITIONS OF TRAINING AND TESTING DATASETS OF THREE SATELLITES

Satellite	Training	Testing (Reduced)	Testing (Full)
GF-2	9000	144	144
WV-2	6000	66	66
WV-3	8000	112	112

from the PAN image and upsampled LRMS image, respectively. All convolutional layers are followed by PReLU activation function. The extracted features F_m and F_p are concatenated as the input of the subsequent DEDB.

The role of DEDB is to learn more high-level features and fuse sufficient spatial and spectral information. As shown in Fig. 2(b), we adopt four convolutional layers with dense connections to enhance the fusion and inference abilities. Then, the fused features are fed into a deconvolutional layer for upsampling before concatenation with the two residual connections. To reconstruct the output HRMS image, we use a reconstruction block (REC) that consists of two convolutional layers followed by a *ReLU* activation layer as demonstrated in Fig. 2(e).

GB and RB are vital parts of our proposed unsupervised model. Taking the output HRMS image or the upsampled LRMS image as the input, GB is implemented aided by the channel attention mechanism, as shown in Fig. 2(c). First, we adopt two convolutional layers to transform the input into weight features and use global average pooling (GAP) and fully connected layers to obtain the channel weight vector, which is used to simulate the graying process. Finally, we obtain the stacked output by copying it in the channel dimension. For RB, we implement this module by using a single convolution layer to simulate the spatial degradation as illustrated in Fig. 2(d). Additionally, these modules are jointly optimized to adaptively learn the degradation in the training phase.

IV. EXPERIMENTS AND EVALUATIONS

A. Experimental Setup

1) Datasets and Metrics: To evaluate the performance of the proposed method, we conduct experiments on three datasets: GaoFen-2 (GF-2), Worldview-2 (WV-2), and Worldview-3 (WV-3). The spatial resolutions of the MS and PAN images for GF-2 satellite are 3.2 m and 0.8 m, respectively, those for WV-2 satellite are 1.84 m and 0.46 m and those for WV-3 satellite are 1.2 m and 0.31 m. The satellite of GF-2 has four bands, while the satellites of later two have eight bands. We produced the training data following the Wald's protocol [19], cropping the PAN and upsampled LRMS images into patch pairs of size 256 × 256 in the training phase. Furthermore, another pairs of size 512 × 512 were selected to implement test experiments of the reduced resolution and full resolution. The partitions of both datasets are listed in Table I.

The performance of different methods in the reduced-resolution and full-resolution experiments are evaluated by different quantitative metrics. In reduced-resolution testing, four widely used metrics with reference are involved, namely, the spectral angle mapper (SAM) [42], spatial correlation coefficient (SCC) [43], relative global synthesis errors (ERGAS) [44], and 4-band extension of the universal image quality index (Q4) [45], while the quality with no-reference (QNR) [46] and its spectral components D_λ and spatial components D_S are used in full-resolution testing.

2) Implementation Details: No postprocessing operations were applied on the output HRMS image. The network was trained with approximately 50 epochs. The Adam optimizer [47] was used to minimize the loss function, with an initial learning rate of 1e−4, and it was decayed by 0.1 every 10 epochs. The batch size was set to 16, the weight of loss α was set to 1, β was set to 5, μ was set to 0.1, δ was set to 20, and γ was set to 20. The network was implemented in PyTorch and trained on an Nvidia GeForce GTX 1080Ti GPU. The codes for this work can be downloaded.[1]

3) Comparison Methods: In our experiments, we compared the proposed LDP-Net with several state-of-the-art methods, including PCA [4], IHS [5], Brovey [48], GS [49], BSBD [11], additive wavelet luminance proportional (AWLP) [50], PNN [14], DiCNN [32], PanNet [15], DMDNet [51], FusionNet [33], PGMAN [36], and Pan-GAN [20]. The first six methods belong to traditional method. PNN, DiCNN, PanNet, DMDNet, and FusionNet are supervised learning based methods. Pan-GAN and PGMAN are recently proposed unsupervised methods. For fair comparison, these methods were reimplemented with the PyTorch framework according to their publicly available codes and retrained using the same training datasets at the reduced resolution.

B. Comparison at Reduced Resolution

The experiment was performed on three datasets at reduced resolution, which follows the Wald's protocol. The original MS image can be used as the reference. Figs. 3–5 show three examples cropped from the results of GF-2, WV-2, and WV-3 processed using different methods. In each case, one region that is marked by a red rectangle is magnified to visualize the differences of these results. In Figs. 3–5, it can be observed that the results of traditional methods can restore spatial details effectively but still exhibit some blurring effects and spectral distortions. For example, the results of BDSD suffer from severe spectral distortions and some blurring effects, while the results of AWLP reduce the blurring effect but introduce some spatial artifacts. Supervised learning based methods can improve the spectral performance of pansharpening results but still exist spatial blurring. For the unsupervised method, Pan-GAN successfully achieves unsupervised pansharpening but its results contain some spatial blurring and obvious spectral distortions, especially in WV-2 and WV-3 datasets. In Fig. 3(n), PGMAN recovers more spatial details in pansharpened results while still exists some spectral distortions. Moreover, GAN-based pansharpening methods are difficult to tune the hyperparameters and easily generate spatial and spectral artifacts. As shown in the magnified regions in Figs. 3(o) and 5(o), compared to other methods, it can be seen that our proposed LDP-Net effectively recovers spatial details and preserves spectral information without introducing artifacts and the fusion results are more vivid and much closer to the ground truth than other methods.

Tables II–IV show the average values of the quantitative results of different methods on three datasets. The methods are classified into three groups, including traditional, supervised, and unsupervised, and the best result in each group are highlighted in bold. Compared with Pan-Gan and PGMAN, the proposed LDP-Net achieves better scores in most metrics. Among the CNN-based methods, the proposed method can

[1] [Online]. Available: https://github.com/suifenglian/LDP-Net

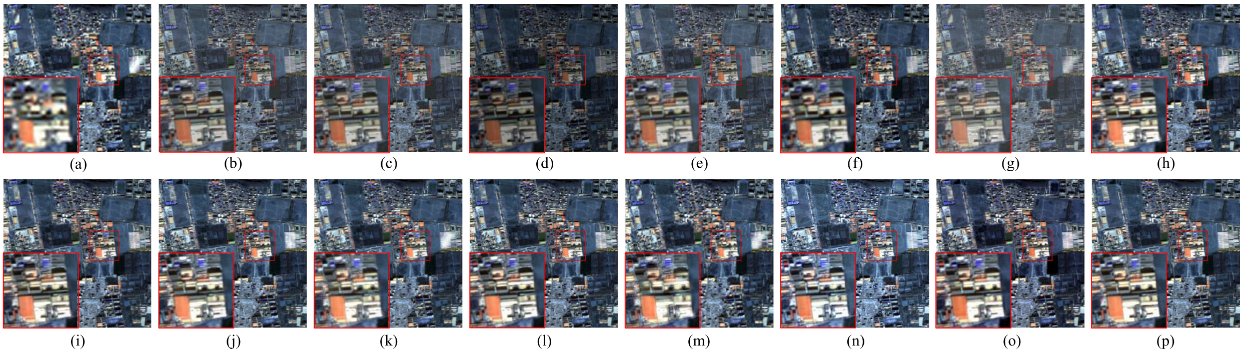


Fig. 3. Pansharpened results from different methods on the GF-2 dataset at reduced resolution. (a) Upsampled LRMS. (b) PCA. (c) IHS. (d) Brovey. (e) GS. (f) BDSD. (g) AWLP. (h) PNN. (i) DiCNN1. (j) PanNet. (k) DMDNet. (l) FusionNet. (m) Pan-GAN. (n) PGMAN. (o) Ours. (p) Ground truth.

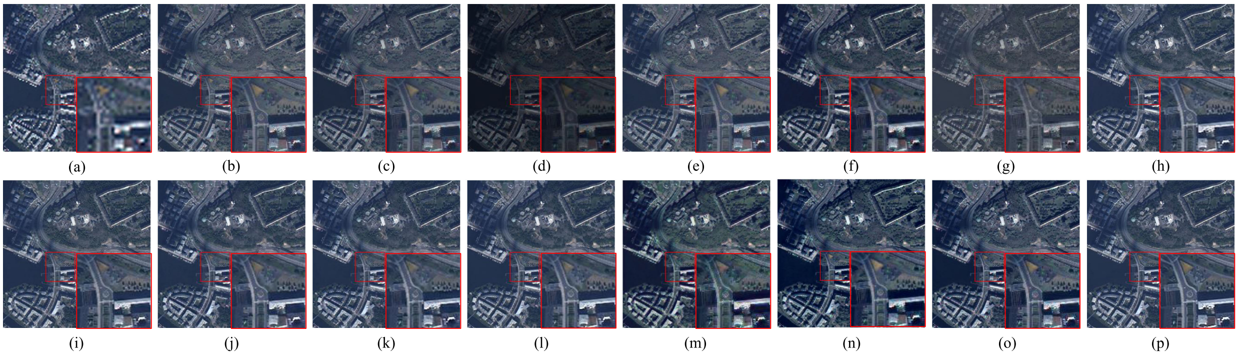


Fig. 4. Pansharpened results from different methods on the WV-2 dataset at reduced resolution. (a) Upsampled LRMS. (b) PCA. (c) IHS. (d) Brovey. (e) GS. (f) BDSD. (g) AWLP. (h) PNN. (i) DiCNN1. (j) PanNet. (k) DMDNet. (l) FusionNet. (m) Pan-GAN. (n) PGMAN. (o) Ours. (p) Ground truth.

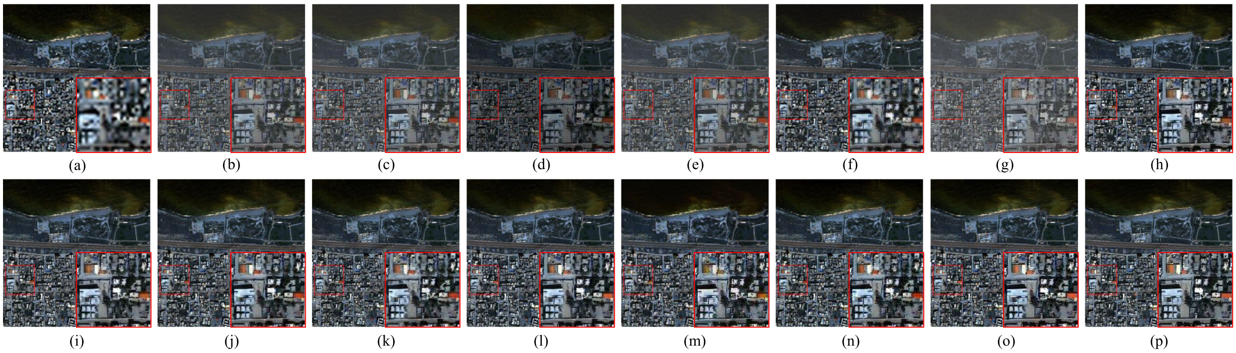


Fig. 5. Pansharpened results from different methods on the WV3 dataset at reduced resolution. (a) Upsampled LRMS. (b) PCA. (c) IHS. (d) Brovey. (e) GS. (f) BDSD. (g) AWLP. (h) PNN. (i) DiCNN1. (j) PanNet. (k) DMDNet. (l) FusionNet. (m) Pan-GAN. (n) PGMAN. (o) Ours. (p) Ground truth.

approach the performance of supervised methods. In particular, our model achieves the SCC, ERGAS, and Q4 scores close to the supervised methods, which verifies that our proposed method can effectively fuse the spatial and spectral information without the reference.

C. Comparison at Full Resolution

In this section, all the methods were validated on real data. Figs. 6–8 illustrate the representative results of the real GF-2, WV-2, and WV-3 data. Moreover, to verify the robustness of the proposed LDP-Net, the models trained with reduced images were used for the full-resolution test, which means we do not need to train new models for the full-resolution datasets. In these cases, most traditional methods can significantly restore the spatial information compared with that in LRMS images but most still suffer from a certain degree of spectral shift. In contrast, AWLP reduces the spectral distortion in the results while introduce noticeable spatial artifacts. Compared with these traditional methods, CNN-based models can effectively maintain spectral consistency and improve the spatial resolution over different datasets. However, PanNet and DMDNet generate perceptible blurring effects and artifacts. DiCNN1 can restore the spatial details better with a high spectral resolution, but spectral

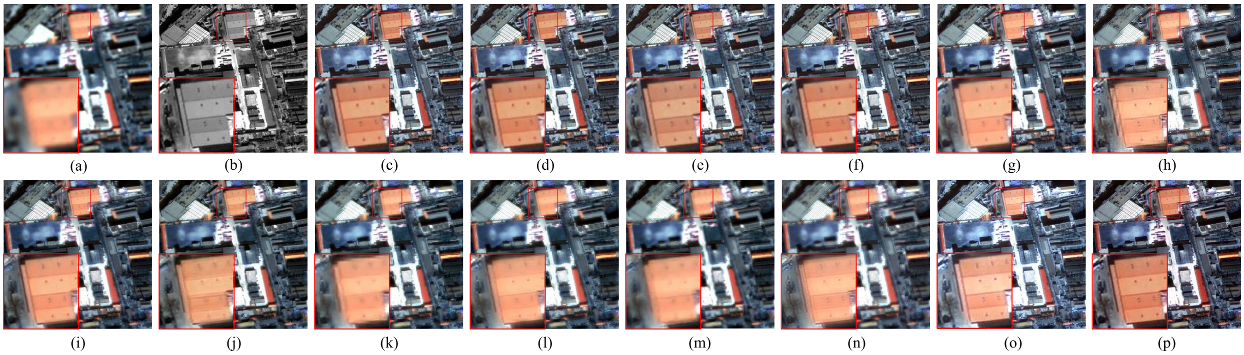


Fig. 6. Pansharpened results from different methods on the GF-2 dataset at full resolution. (a) Upsampled LRMS. (b) PAN. (c) PCA. (d) IHS. (e) Brovey. (f) GS. (g) BDSD. (h) AWLP. (i) PNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) FusionNet. (n) Pan-GAN. (o) PGMAN. (p) Ours.

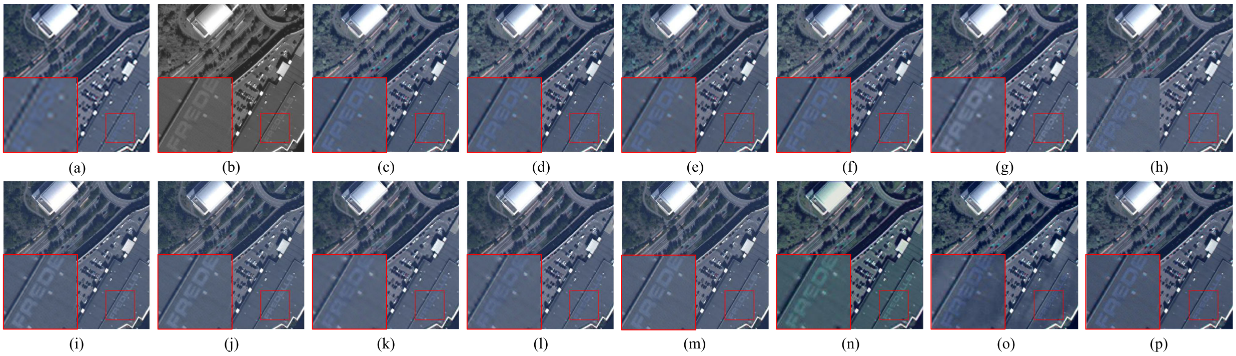


Fig. 7. Pansharpened results from different methods on the WV-2 dataset at full resolution. (a) Upsampled LRMS. (b) PAN. (c) PCA. (d) IHS. (e) Brovey. (f) GS. (g) BDSD. (h) AWLP. (i) PNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) FusionNet. (n) Pan-GAN. (o) PGMAN. (p) Ours.

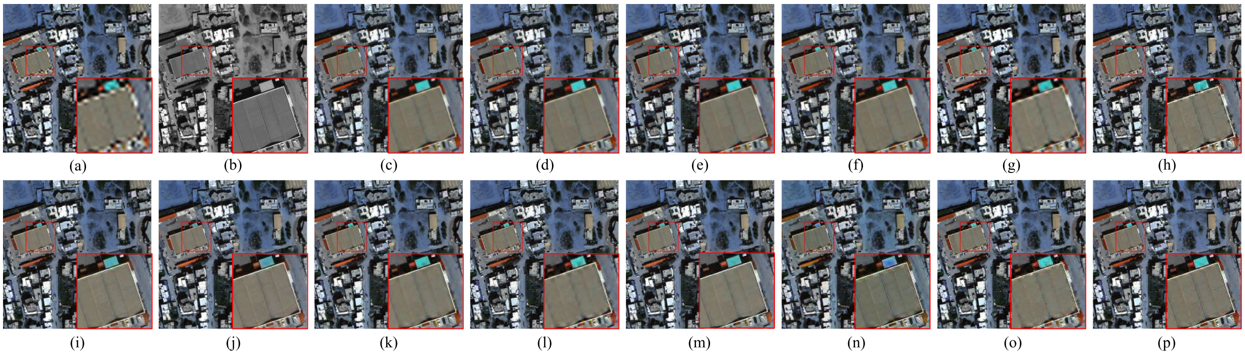
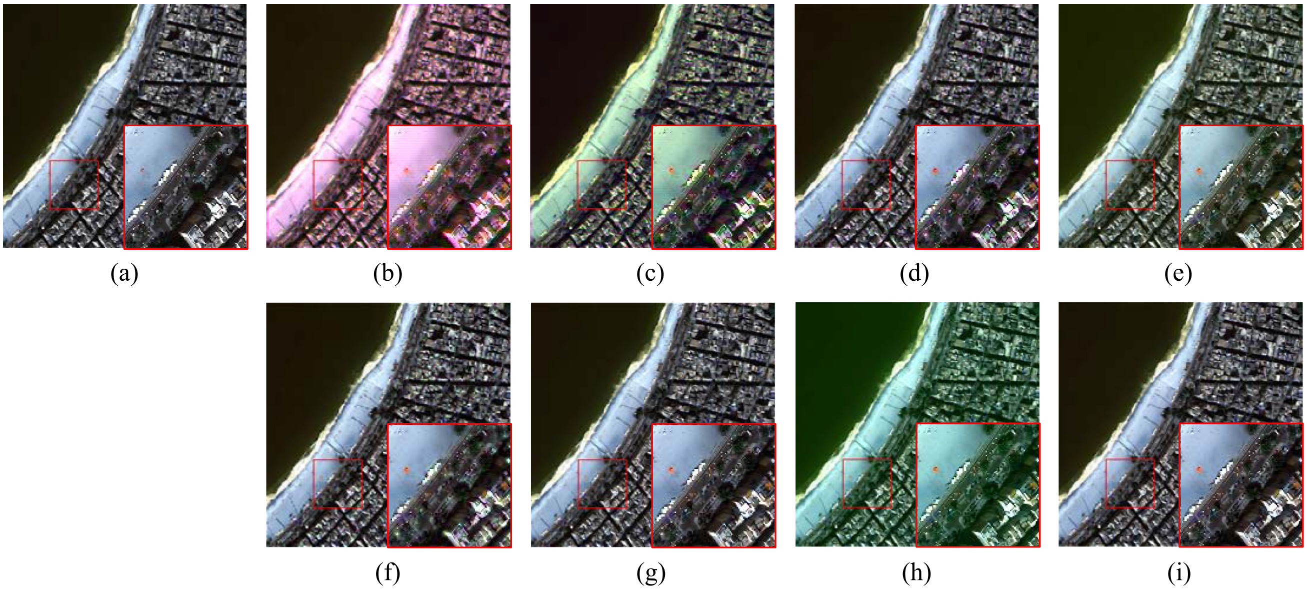


Fig. 8. Pansharpened results from different methods on the WV-3 dataset at full resolution. (a) Upsampled LRMS. (b) PAN. (c) PCA. (d) IHS. (e) Brovey. (f) GS. (g) BDSD. (h) AWLP. (i) PNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) FusionNet. (n) Pan-GAN. (o) PGMAN. (p) Ours.

distortions are still observed in parts of regions. As shown in Figs. 7(j) and 8(j), the light blue mark and the cyan buildings are not as vividly colored as those obtained by other methods. Compared with other supervised methods, FusionNet can further reduce the spatial blurring and spectral distortions. Pan-GAN, which achieves unsupervised learning using spatial and spectral discriminators, can improve the spatial and spectral resolution but still exist spatial blurring and introduce spectral distortions to the results in Figs. 7(n) and 8(n). Though PGMAN maintains the spectral consistency as the upsampled LRMS image, there are still noticeable distortions of spatial details in pansharpened results. It is obvious that in the magnified regions indicated by red boxes, our proposed method preserves better spatial details and maintains higher spectral consistency than other methods. Apparently, our pansharpened images are clearer and more vivid than all the other methods, as shown in Figs. 6(p) and 8(p).

Due to lack of ground truth, QNR, D_λ, and D_S are employed as the quantitative metrics to evaluate the performance of the pansharpened results at full resolution. The quantitative results are shown in Table V. As shown in Table V, we notice that the quantitative results are not quite consistent with the results of the visual inspections. This paradox probably lies in that


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

(f) (g) (h) (i)

Fig. 9. Pansharpened results from the ablation study of the loss functions. (a) Ground truth. (b) The combination I. (c) The combination II. (d) The combination III. (e) The combination IV. (f) The combination V. (g) The combination VI. (h) The combination VII. (i) The combination VIII.

TABLE II
QUANTITATIVE RESULTS ON THE GF-2 DATASET AT REDUCED RESOLUTION

Type	Method	SAM↓	SCC↑	ERGAS↓	Q4↑
Traditional	PCA	6.9380	0.9052	**2.2390**	**0.9655**
	IHS	7.5092	0.9017	2.3679	0.9630
	Brovey	7.1863	0.8972	2.4259	0.8686
	GS	7.4763	0.9031	2.3447	0.9595
	BDSD	**5.7480**	0.8655	2.4987	0.8319
	AWLP	7.8057	**0.9295**	2.4278	0.8779
Supervised	PNN	4.4652	0.9292	3.3447	0.9884
	DiCNN1	4.2337	0.9446	3.1716	0.9980
	PanNet	4.1659	0.9379	3.2079	0.9996
	DMDNet	4.1237	0.9619	2.5558	**0.9998**
	FusionNet	**3.7556**	**0.9662**	**2.4327**	0.9992
Unsupervised	Pan-GAN	**4.6391**	0.8918	4.7974	0.9890
	PGMAN	9.2761	0.9378	5.3594	0.9706
	Ours	6.3531	**0.9676**	**3.3210**	**0.9895**

TABLE III
QUANTITATIVE RESULTS ON THE WV-2 DATASET AT REDUCED RESOLUTION

Type	Method	SAM↓	SCC↑	ERGAS↓	Q8↑
Traditional	PCA	4.3250	**0.8669**	2.4778	0.9573
	IHS	4.0705	0.8572	**2.4521**	0.9636
	Brovey	4.3178	0.8198	2.5145	0.8984
	GS	4.3531	0.8659	2.4982	0.9556
	BDSD	4.1415	0.8271	2.8372	0.9145
	AWLP	**3.9694**	0.8371	2.5660	**0.9772**
Supervised	PNN	3.2113	0.8790	3.6128	0.9884
	DiCNN1	3.1771	0.8894	3.5280	0.9971
	PanNet	3.0514	0.9000	3.4314	0.9988
	DMDNet	3.0137	0.9082	3.1872	0.9986
	FusionNet	**2.8311**	**0.9165**	**3.0281**	**0.9989**
Unsupervised	Pan-GAN	6.4173	0.8108	5.7978	0.9800
	PGMAN	5.6582	0.8401	5.3877	0.9710
	Ours	**4.3543**	**0.8567**	**5.0884**	**0.9821**

TABLE IV
QUANTITATIVE RESULTS ON THE WV-3 DATASET AT REDUCED RESOLUTION

Type	Method	SAM↓	SCC↑	ERGAS↓	Q8↑
Traditional	PCA	8.3785	0.8548	4.1581	0.7784
	IHS	7.8421	**0.8601**	4.0947	**0.7830**
	Brovey	**7.3914**	0.8447	4.1840	0.7392
	GS	8.1898	0.8567	**4.0663**	0.7667
	BDSD	8.0150	0.7893	4.6131	0.7830
	AWLP	10.7938	0.8277	5.0647	0.7569
Supervised	PNN	7.6609	0.8637	8.0054	0.8841
	DiCNN1	7.1150	0.8886	7.4866	0.9051
	PanNet	7.4844	0.8709	7.7732	0.8973
	DMDNet	6.8661	0.8937	6.9935	0.9803
	FusionNet	**6.4131**	**0.8975**	**6.9463**	**0.9835**
Unsupervised	Pan-GAN	10.0138	0.8464	9.5112	0.8510
	PGMAN	8.6667	0.8647	6.6335	0.9151
	Ours	**7.8846**	**0.9347**	**4.999**	**0.9574**

the nonreference assessment metrics are calculated using the LRMS images, PAN image, and pansharpened results to assess the spectral and spatial distortion. The results with blurring effects tend to achieve better values due to their similarity to the LRMS images, which has also been mentioned in [52]. For example, as shown in Figs. 7(j) and (l), and 8(j) and (l), the results of DMDNet with more obvious blurring effects have better QNR values than DiCNN1. Hence, nonreference metrics are not always suitable to assess the spectral and spatial distortions of pansharpened results, and it is more important to emphasize visual inspection for comparison at full resolution without ground truth.

D. Ablation Study of Loss Function

In this section, several experiments were conducted to investigate the impacts of each component in our loss function. Based on two learnable degradation processes, the loss function plays an important role in our unsupervised training process. The proposed loss function can be subdivided into five parts, namely, the spatial loss at high resolution $L_{spatial\_h} = \|\widetilde{P} - \widehat{M}_{gray}\|_2^2$, the spatial loss at low resolution $L_{spatial\_l} = \|\widetilde{P}_{blur} - \uparrow m_{gray}\|_2^2$, the spectral loss at high resolution $L_{spectral\_h} = \|\uparrow m - \widehat{M}_{blur}\|_2^2$, the spectral loss at low

TABLE V
QUANTITATIVE RESULTS AT FULL RESOLUTION

		GF-2			WV-2			WV-3		
Type	Method	$D_\lambda \downarrow$	$D_S \downarrow$	QNR↑	$D_\lambda \downarrow$	$D_S \downarrow$	QNR↑	$D_\lambda \downarrow$	$D_S \downarrow$	QNR↑
Traditional	PCA	0.0096	0.0900	0.9013	0.0115	0.1038	0.8864	0.0083	0.0615	0.9309
	IHS	0.0066	0.0926	0.9015	**0.0030**	0.1101	0.8875	0.0082	0.0666	0.9266
	Brovey	0.0285	0.0825	0.8914	0.0173	0.1063	0.8790	0.0119	0.0605	0.9285
	GS	**0.0093**	0.0924	0.8993	0.0122	0.1052	0.8843	**0.0042**	0.0621	**0.9340**
	BDSD	0.1781	0.1731	0.7274	0.0116	0.0679	**0.9225**	0.0227	**0.0557**	0.9247
	AWLP	0.0246	**0.0635**	**0.9136**	0.0310	**0.0616**	0.9131	0.0597	0.0901	0.8658
Supervised	PNN	0.0124	0.0284	0.9596	0.0418	0.0475	0.9142	0.0348	0.0896	0.8788
	DiCNN1	0.0165	**0.0218**	**0.9621**	0.0501	0.0612	0.8958	0.0319	0.0917	0.8805
	PANNET	0.0086	0.0611	0.9308	0.0337	0.0322	0.9342	**0.0236**	0.0823	0.8967
	DMDNet	0.0084	0.0341	0.9578	0.0264	0.0321	0.9423	0.0241	0.0782	0.9001
	FusionNet	**0.0022**	0.0404	0.9574	**0.0221**	**0.0316**	**0.9470**	**0.0207**	**0.0761**	**0.9048**
Unsupervised	Pan-GAN	**0.0073**	0.1965	0.7976	0.0891	0.1776	0.7491	0.0379	0.1904	0.7789
	PGMAN	0.0276	0.1001	0.8751	**0.0425**	0.1538	0.8193	0.0312	0.1116	0.8607
	Ours	0.0263	**0.0321**	**0.9427**	0.0437	**0.1477**	**0.8232**	**0.0287**	**0.0971**	**0.8774**

TABLE VI
ABLATION RESULTS WITH THE LOSS FUNCTIONS ON WV-3 DATASET

The combination of losses	Losses			SAM	SCC	ERGAS	Q8
	$L_{spatial\_l}$	$L_{spectral\_l}$	L_{KL}				
I				27.7327	0.8719	38.3513	0.7206
II	√			20.7888	0.9074	11.2121	0.8239
III		√		10.5851	0.8734	8.4072	0.9129
IV			√	14.6685	0.9374	18.0237	0.8616
V	√	√		9.5221	0.8929	7.0742	0.8728
VI		√	√	8.1887	0.9264	5.1832	0.9473
VII	√		√	21.2254	0.9310	79.2213	0.8102
VIII	√	√	√	**7.8846**	**0.9347**	**4.9991**	**0.9574**

resolution $L_{spectral\_l} = \|m - \downarrow \widehat{M}\|_2^2$ and the spectral KL divergence loss L_{KL}. $L_{spatial\_h}$ and $L_{spectral\_h}$ are used as the basic loss components for the unsupervised training. Table VI shows the quantitative results to validate the effectiveness of $L_{spatial\_l}$, $L_{spectral\_l}$ and the proposed spectral KL divergence loss. In addition, we display the visual results of different combinations of loss components in Fig. 9. It can be seen that the combination of only $L_{spatial\_h}$ and $L_{spectral\_h}$ cannot achieve satisfactory performance, which suffers from severe spectral distortions in pansharpened images. Low-resolution spatial loss can restore the spatial details but still suffer from spectral distortions, and low-resolution spectral loss can reduce the spectral distortions but produce some spectral artifacts, while the spectral KL divergence loss can obviously eliminate spectral artifacts with high spatial resolution but still remain spectral distortions. When all of the loss components are included, the pansharpened images have the best quantitative scores and achieve the best spatial and spectral consistency, fully utilizing the rich spatial information of HR PAN images and the relation between MS images and PAN images. These results verify the effectiveness of our proposed hybrid loss function in both qualitative and quantitative aspects.

E. Efficiency Study

In this section, the computational efficiencies of all comparison methods are evaluated. As mentioned in Section IV-A, all deep learning based methods were implemented in PyTorch and tested on an Nvidia GeForce GTX 1080Ti GPU, while all traditional methods were implemented in MATLAB R2019b framework on CPU. Table VII lists the computational times of different approaches and the parameters of different models. The cost times are evaluated by averaging the inference time in the testing set at the reduced resolution experiment. Compared with other methods, the number of the parameters of our model is small but the computational time of our method is at the middle level. The main reason is that our proposed network contains two additional degradation modules and a deeper network structure. Compared to GAN-based unsupervised pansharpening methods, we must mention that our model is easier for hyperparameter tuning in the training phase. Generally, in addition to ensuring the superiority of performance, our proposed unsupervised model makes a reasonable tradeoff between model performance and computational cost.

TABLE VII
EFFICIENCY COMPARISON WITH DIFFERENT METHODS WHEN PROCESSING INPUTS OF SIZE 256 × 256 × 4 ON GF-2 DATASET

Type	Method	Run time(s)	Params(M)
Traditional	PCA	0.0729	–
	IHS	0.0080	–
	Brovey	0.0079	–
	GS	0.0902	–
	BDSD	0.3593	–
	AWLP	0.2340	–
Supervised	PNN	0.0167	0.080
	DiCNN1	0.0188	0.080
	PanNet	0.0158	0.078
	DMDNet	0.0149	0.029
	FusionNet	0.0163	0.076
Unsupervised	Pan-GAN	0.0129	0.092
	PGMAN	0.0348	3.914
	Ours	0.0346	0.071

V. CONCLUSION

In this article, we propose an unsupervised pansharpening method based on two learnable degradation processes. The method can adaptively learn the degraded processes with two corresponding CNN-based modules and successfully achieve unsupervised pansharpening. Moreover, we consider the degradation processes at different resolutions and present a novel hybrid loss that can effectively maintain spatial and spectral consistency. Thus, this unsupervised training strategy adequately improves the spatial details and reduces the spectral distortion in the results. Then, extensive experiments were performed on different-resolution images from three datasets, demonstrating the superiority of our proposed method over other state-of-the-art methods.

References

[1] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, "Deep learning-based fusion of landsat-8 and sentinel-2 images for a harmonized surface reflectance product," *Remote Sens. Environ.*, vol. 235, 2019, Art. no. 111425.

[2] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[3] R. Haydn, G. W. Dalke, J. Henkel, and J. C. Bare, "Application of the IHS color transform to the processing of multisensor data and image enhancement," in *Proc. Int. Symp. Remote Sens. Arid Semi-Arid Lands*, Cairo, Egypt, 1982, pp. 599–607.

[4] 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.

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

[6] 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.

[7] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[8] P. S. Pradhan, R. L. King, N. H. Younan, and D. W. Holcomb, "Estimation of the number of decomposition levels for a wavelet-based multiresolution multisensor image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3674–3686, Dec. 2006.

[9] S. Zheng, W.-Z. Shi, J. Liu, G.-X. Zhu, and J.-W. Tian, "Multisource image fusion method using support value transform," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1831–1839, Jul. 2007.

[10] 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, 2006.

[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. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.

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

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

[15] 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.*, 2017, pp. 5449–5457.

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

[17] 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.

[18] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, 2020.

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

[20] 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, 2020.

[21] 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, pp. 1–16, 2016.

[22] P. Chavez *et al.*, "Comparison of three different methods to merge multiresolution and multispectral data- landsat TM and spot panchromatic," *Photogrammetric Eng. Remote Sens.*, vol. 57, no. 3, pp. 295–303, 1991.

[23] W. Dou, Y. Chen, X. Li, and D. Z. Sui, "A general framework for component substitution image fusion: An implementation using the fast image fusion method," *Comput. Geosci.*, vol. 33, no. 2, pp. 219–228, 2007.

[24] 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.

[25] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[26] W. Wright, "Fast image fusion with a Markov random field," in *Proc. 7th Int. Conf. Image Process. Appl.*, 1999, pp. 557–561.

[27] M. Guo, H. Zhang, J. Li, L. Zhang, and H. Shen, "An online coupled dictionary learning approach for remote sensing image fusion," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1284–1294, Apr. 2014.

[28] P. Guo, P. Zhuang, and Y. Guo, "Bayesian pan-sharpening with multiorder gradient-based deep network constraints," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 950–962, Mar. 2020.

[29] 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.

[30] 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.

[31] 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.

[32] L. He *et al.*, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.

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

[34] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, and G. Vivone, "A triple-double convolutional neural network for panchromatic sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, 2022, doi: 10.1109/TNNLS.2022.3155655.

[35] C. Jin, L.-J. Deng, T.-Z. Huang, and G. Vivone, "Laplacian pyramid networks: A new approach for multispectral pansharpening," *Inf. Fusion*, vol. 78, pp. 158–170, 2022.

[36] H. Zhou, Q. Liu, and Y. Wang, "PGMAN: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6316–6327, Mar. 2022.

[37] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 87–102.

[38] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, Jul. 2020.

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

[40] A. Garzelli, "A review of image fusion algorithms based on the super-resolution paradigm," *Remote Sens.*, vol. 8, no. 10, Art. no. 797, 2016.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[42] 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, vol. 1, pp. 147–149.

[43] J. Zhou, D. L. Civco, and J. 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.

[44] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *Proc. 3rd Conf. Fusion Earth Data: Merging Point Meas., Raster Maps Remotely Sensed Images*, 2000, pp. 99–103.

[45] 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.

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

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[48] 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, 1987.

[49] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6,011,875, Jan. 4, 2000.

[50] Y. Kim, C. Lee, D. Han, Y. Kim, and Y. Kim, "Improved additive-wavelet image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 263–267, Mar. 2011.

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

[52] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3192–3208, Apr. 2021.

Jiahui Ni received the M.S. degree in communications and information system from the College of Electronics and Information Engineering, Sichuan University, Chengdu, China, in 2022.

His research interests include deep learning, computer vision and hyperspectral image processing.

Zhimin Shao received the M.S degree in computer science and technology from the College of Computer Science, Sichuan University, Chengdu, China, in 2020.

His main research interests in computer version, image super-resolution and image denosing, etc.

Zhongzhou Zhang is currently working toward the Ph.D degree with the College of Computer Science, Sichuan University, Chengdu, China. He received the B.S. and M.S. degrees in computer science and technology from the Tianjin University of Technology and Chongqing University, respectively.

His main research interests including medical image analysis, domain generalization, etc.

Mingzheng Hou received the B.S., M.S., and Ph.D. degrees in computer science and technology from the College of Computer Science, Sichuan University, Chengdu, China, in 2010, 2013, and 2022, respectively.

She is currently with the College of Computer Science, Sichuan University, Chengdu. Her main research interests include computer vision, image superresolution, and target detection, etc.

Jiliu Zhou (Senior Member, IEEE) received the Ph.D. degree in hydrology and water resources from the College of Water Resource & Hydropower, Sichuan University, Chengdu, China, in 1999.

He is currently a Professor with the School of Computer Science, Sichuan University, Chengdu, China. He is also the Principal of Chengdu University of Information Technology, Chengdu. His research is mainly in the field of image processing, artificial intelligence, fractional differential application on the latest signal and image processing, etc. He has published more than 200 papers, of which more than 80 papers are indexed by SCI, EI, or ISTP.

Leyuan Fang (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015. From September 2011 to September 2012, he was a Visiting Ph.D. Student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council.

From August 2016 to September 2017, he was a Postdoctoral Researcher with the Department of Biomedical Engineering, Duke University. He is a Professor with the College of Electrical and Information Engineering, Hunan University, and an Adjunct Researcher with the Peng Cheng Laboratory, Shenzhen, China. His research interests include sparse representation and multiresolution analysis in remote sensing and medical image processing.

Dr. Fang was a recipient of one 2nd-Grade National Award at the Nature and Science Progress of China in 2019. He is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and NEUROCOMPUTING.

Yi Zhang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and technology from the College of Computer Science, Sichuan University, Chengdu, China, in 2005, 2008, and 2012, respectively.

From 2014 to 2015, he was with the Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA, as a Postdoctoral Researcher. He is currently a Full Professor with the School of Cyber Science and Engineering, Sichuan University, and is the Director of the deep imaging group (DIG). His research interests include medical imaging, compressive sensing, and deep learning. He authored more than 80 papers in the field of image processing. These papers were published in several leading journals, including IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING, *Medical Image Analysis, European Radiology, Optics Express*, etc., and reported by the Institute of Physics (IOP) and during the Lindau Nobel Laureate Meeting. He received major funding from the National Key R&D Program of China, the National Natural Science Foundation of China, and the Science and Technology Support Project of Sichuan Province, China. He is a Guest Editor of the *International Journal of Biomedical Imaging*, *Sensing and Imaging*, and an Associate Editor of IEEE ACCESS.