MSSL: Hyperspectral and Panchromatic Images Fusion via Multiresolution Spatial–Spectral Feature Learning Networks

Jiahui Qu, Member, IEEE, Yanzi Shi, Student Member, IEEE, Weiying Xie, Member, IEEE, Yunsong Li, Member, IEEE, Xianyun Wu, and Qian Du, Fellow, IEEE

Abstract—The fusion of hyperspectral (HS) and panchromatic (PAN) images aims to generate a fused HS image that combines spectral information of the HS image with spatial information of the PAN image. In this article, we propose a multiresolution spatial–spectral feature learning (MSSL) framework for fusing HS and PAN images. The proposed MSSL transforms the existing deep and complex network into several simple and shallow subnetworks to simplify the feature learning process. MSSL upsamples the HS image while downsamples the PAN image and designs multiresolution 3-D convolutional autoencoder (CAEs) networks with a spectral constraint to learn complete spatial–spectral features of the HS image. MSSL designs multiresolution 2-D CAEs with spatial constraint to extract spatial features of the PAN image, with a low computational cost. In order to effectively generate the pansharpened HS image with high spatial and spectral fidelity, a multiresolution residual network is presented to reconstruct the HS image from the extracted spatial–spectral features. Extensive experiments are conducted on three widely used remote sensing data sets in comparison with state-of-the-art HS image fusion methods, demonstrating the superiority of the proposed MSSL method. Code is available at https://github.com/JiahuiQu/MSSL.

Index Terms—Convolutional autoencoder (CAE), hyperspectral (HS) pansharpening, image fusion, multiresolution, spatial–spectral feature.

I. INTRODUCTION

WITH the development of remote sensing technology, hyperspectral (HS) images with dozens or even hundreds of spectral bands can be acquired. The continuous and dense sampling in a given spectral range makes it possible to distinguish similar materials. However, for an optical remote sensing system with a given signal-to-noise ratio (SNR), the high spectral resolution of HS image is achieved at the cost of spatial resolution. The fusion of HS and panchromatic (PAN) images is a powerful image enhancement method that combines spectral information of low-resolution (LR) HS images with spatial information of high-resolution (HR) PAN images of the same scene to generate HS images with high spatial and spectral resolutions. It constitutes the possibility to obtain HR images in both spectral and spatial domains, which is difficult for a single sensor due to physical constraints. The HS and PAN images fusion, also called HS pansharpening, can greatly improve image interpretation ability in practical applications, such as change detection [1], object recognition [2], anomaly detection [3], [4], and classification [5], [6]. There has been a lively interest since pansharpening in the last few decades.

A series of HS pansharpening models have been put forward in the literature. They can be classified into the following four categories: component substitution (CS), multisolution analysis (MRA), Bayesian, and matrix factorization. The CS-based methods rely on replacing the spatial component of HS images with the histogram-matched PAN images. Among so many CS-based algorithms, the intensity–hue–saturation (IHS) [7], [8] method, the principal component analysis (PCA) [9]–[11] method, the Gram–Schmidt (GS) transform-based method [12], and the adaptive GS (GSA) [13] method are the best representatives. Although this class is, in general, easy to implement, spectral distortion may be caused due to a mismatch of spectral range between PAN and HS data [14]. In order to solve this problem, a number of methods are proposed. Dong et al. [15] improved the CS method and proposed a novel detail extraction and injection gain estimation method. The MRA-based methods extract spatial details from PAN images and inject them into the interpolated HS images, which has the advantage of spectral consistency. This class relies on multiscale image decomposition techniques, for example, decimated wavelet transforms (DWTs) [16], undecimated wavelet transform (UDWT) [17], a trous wavelet transform (ATWT) [18], and Laplacian pyramid (LP) [19]. The most representative methods of the MRA class are smoothing filter-based intensity modulation (SFIM) [20], modulation transfer function generalized...
LP (MTF-GLP) [21], MTF-GLP with high pass modulation (MGH) [22], and MTF-GLP with context-based decision (MTF-GLP-CBD) [23]. They are generally superior to the CS class in preserving spectral contents but suffer from spatial distortion. The Bayesian methods use prior information to regularize the ill-posedness problem of HS pansharpening. Related works include the Bayesian sparsity promoted Gaussian prior (BSF) [24] method, the Bayesian HySure [25] method, and the Bayesian naive Gaussian prior (Bayesian naive) [26] method. Matrix factorization is also an effective method to deal with HS pansharpening. For example, the algorithm proposed by Yokoya et al. [27] alternately unmixed HS and MS data to obtain the endmember spectra and HR abundance maps. The Bayesian and matrix factorization methods perform well in terms of the preservation of spectral information, but they have a large computational amount. Most of the traditional HS pansharpening methods were originally designed and developed for multispectral pansharpening. However, there is a large spectral mismatch range between HS and PAN images. Therefore, how to produce a good fusion performance in the spectral range that is not covered by PAN images is a key issue to be broken through.

In the last few years, convolutional neural networks (CNNs) with high nonlinearity are very effective in exploiting image characteristics and have achieved remarkable success in image processing fields, such as super-resolution [28], target detection [29], image segmentation [30], and image classification [31]. Dong et al. [32] successfully proposed an image super-resolution CNN (SRCNN) by adopting the CNN architecture to learn the mapping between LR-HS and HR-HS images and achieved a remarkable super-resolution performance. Inspired by the outstanding performance of deep learning in super-resolution, we intend to solve the pansharpening problem by adopting a deep learning perspective. In recent years, CNN-based methods have become the new trend in pansharpening. Masi et al. [33] proposed an SRCNN-based pansharpening CNN (PNN) with the pre-interpolated LRMS and the PAN image as inputs to learn the mapping between the stacked input image and the output HRMS image. Wei et al. [34] proposed a deep residual-based pansharpening method (DRPNN), which improves its performance, due to the deeper network architecture. However, most of the existing CNN-based pansharpening methods are solely based on super-resolution, without physical interpretation of pansharpening. Some methods, such as [35], directly adopt the 3-D convolutional layer designed for super-resolution to implement pansharpening. They cannot make full use of the depth of the network to extract deep features and cannot fully extract the spatial information of the PAN image. In addition, these methods overlay the PAN image as a band on HS image for network training. Since all the bands are trained indiscriminately, the spatial information of the PAN image cannot be fully extracted even if the network depth is increased. Moreover, the deeper the network, the more storage space is required.

In this article, instead of designing a deep and complex network structure, we design simple and shallow multisolution networks for each resolution to simplify the feature learning process. In other words, we propose a multiresolution spatial–spectral feature learning (MSSL) network for HS pansharpening, and the proposed method transforms a difficult deep pansharpening problem into a few simple shallower pansharpening subproblems. More specifically, the proposed MSSL consists of two main parts: spatial–spectral feature extraction and spatial–spectral feature-based HS reconstruction. The spatial–spectral feature extraction part is designed for extracting complete spatial–spectral features from PAN and HS images, while the HS reconstruction part fuses the extracted spatial–spectral features from different resolutions to yield final pansharpened results. Compared with the traditional classical methods and the existing CNN-based networks, the proposed MSSL has the following contributions.

1) In order to reduce computation cost and extract complete features, the HS image that has a different spatial resolution from the PAN image is upsampled, while the PAN image is downsampled in spatial–spectral feature extraction networks. Several shallow multiresolution networks are designed to extract adequate spatial–spectral features of the obtained multiresolution HS and PAN images, with a low computational cost.

2) In spatial–spectral feature extraction part, we design multiresolution 3-D convolutional autoencoder CAE networks to learn both spatial information and spectral information of the HS image simultaneously and design multiresolution 2-D CAE networks to extract the spatial features of the PAN image.

3) In spatial–spectral feature-based HS reconstruction part, a multiresolution residual reconstruction network is presented to reconstruct the pansharpened HS image with high spatial and spectral resolutions.

4) The spectral constraint and spatial constraint are, respectively, exploited to facilitate multiresolution 3-D CAE networks and multiresolution 2-D CAE networks more accurate and effective.

The rest of this article is organized as follows. The proposed HS pansharpening method is presented in Section II. Section III reports and discusses the experimental results. Finally, Section IV draws the conclusion.

II. PROPOSED METHOD

In this section, the proposed MSSL is presented in detail. Since the HS image and the PAN image have different spatial resolutions, most HS pansharpening methods first interpolate the HS image to the scale of the PAN image and then feed the interpolated HS image into the deep learning network. However, these methods may lose a portion of the spectral features of the original LR HS image. In addition, these methods generally need a deep and complex network to extract more adequate features. In order to solve these problems, the proposed MSSL designs several shallow multiresolution networks to extract adequate spatial–spectral features, without losing spectral information of the original HS image. Based on the spatial–spectral features, a multiresolution residual network is proposed to reconstruct the pansharpened HS image. In specific, the MSSL method mainly includes three parts:
multiresolution 3-D HS spatial–spectral feature extraction network, multiresolution 2-D PAN spatial feature extraction network, and spatial–spectral feature-based multiresolution residual HS reconstruction network. The architecture of the proposed MSSL is illustrated in Fig. 1. Let $H \in \mathbb{R}^{m \times n \times c}$ denote the LR-HS image and $P \in \mathbb{R}^{M \times N}$ denote the HR-PAN image, where $M \times N = (m \times n) \times d^2$ is the number of pixels in the PAN image, $m \times n$ is the number of pixels in one band of the HS image, $d$ denotes the scale factor, and $c$ is the number of spectral bands. Let the desired HR-HS image be denoted as $H_F \in \mathbb{R}^{M \times N \times c}$.

The HS image is a 3-D data cube with two spatial dimensions and one spectral dimension. The PAN image is a 2-D data, which contains high spatial resolution and abundant spatial context. Furthermore, the HS image has a different spatial resolution from the PAN image. In order to well explore adequate features of both HS and PAN images and reduce computation cost, we upsample the HS image while downsampling the PAN image in the network. The obtained multiresolution HS and PAN images are fed into the shallow multiresolution networks to extract spatial–spectral features. Specifically, the multiresolution 3-D CAE networks are designed to extract both spatial information and spectral discrimination of the HS image simultaneously, and the multiresolution 2-D CAE networks are applied to the multiresolution PAN images to capture abundant spatial features. More details of each part are described as follows.

A. Multiresolution 3-D HS spatial–spectral Feature Extraction Network

The proposed MSSL is suitable for the case where the ratio $d$ is $2^i$. This article takes a four-time ratio as an example to describe in detail. As shown in Fig. 1, the proposed MSSL can be divided into two main parts: spatial–spectral feature extraction and spatial–spectral feature-based HS reconstruction. In the spatial–spectral feature extraction part, to effectively process HS and PAN images with different resolutions, we analyze the characteristics of the HS and PAN data. For the sake of acquiring complete features of both HS and PAN images, $2(i + 1)$ shallow multiresolution CAE networks are designed, in which $(i + 1)$ 3-D CAE networks are utilized to extract the spatial–spectral features of 3-D HS images and $(i + 1)$ 2-D CAE networks are applied to the PAN images for spatial feature extraction.

AE [36] is an unsupervised neural network. It can learn a hidden representation of the input data from the unlabeled data through reconstructing the input data with a minimum error.
CAE has the same structure as AE but uses convolutional operations. CAE is more suitable for solving the pansharpening problem than AE, which uses fully connection layers, since convolution considers the structure information. 2-D CAE can automatically learn effective spatial features, while 3-D CAE automatically learns spatial and spectral features simultaneously.

In the MSSL method, taking four times ratio as an example, upsampling the LR-HS image twice can obtain the double-interpolated HS image and the quadruple-interpolated HS image. The shallow 3-D CAEs are applied to three HS images to learn complete spatial-spectral features of the HS data efficiently, in which 3-D convolution explores spectral information and spatial context simultaneously. For each resolution encoder network, the HS image is fed into the 3-D convolution encoder network and is passed through two $3 \times 3 \times 3$ 3-D convolution layers to obtain the hidden feature maps. Rectified linear units (ReLUs) are employed for non-linearity in the multiresolution 3-D CAEs. The obtained hidden representation of each resolution denotes spatial-spectral features of each resolution HS image and can be represented by the following mapping:

$$F_{ss}^{i} = \left\{ \begin{array}{ll}
    f \{ \omega_{ss}^{i} \in [ f \{ \omega_{s1}^{i} \ast H + \beta_{s1}^{i} \} \} + \beta_{ss}^{i}, & i = 0 \\
    f \{ \omega_{ss}^{i} \in [ f \{ \omega_{s1}^{i} \ast (\uparrow H) + \beta_{s1}^{i} \} \} + \beta_{ss}^{i}, & i = 1 \\
    f \{ \omega_{ss}^{i} \in [ f \{ \omega_{s1}^{i} \ast (\downarrow (\uparrow H)) + \beta_{s1}^{i} \} \} + \beta_{ss}^{i}, & i = 2
\end{array} \right. $$

where $F_{ss}^{i}$ is the spatial-spectral features of each resolution HS image, $\uparrow$ represents the upsampling operation, $\omega_{ss}^{i}$ and $\omega_{s1}^{i}$ are the encoder weights of the $i$th resolution, $\beta_{s1}^{i}$ and $\beta_{ss}^{i}$ are the corresponding encoder bias, $f$ represents the ReLU activation function, and $*$ denotes the 3-D convolution. The 3-D convolution of the $j$th convolutional kernel in the $l$th convolution layer is calculated by

$$I_{l,j}^{i,x,y,z} = \sum_{k=1}^{K_{l-1}} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} W_{l,j,k}^{i,p,q,r} I_{l-1,j,k}^{i,x+y+q,z+r} + \beta_{l,j} $$

where $I_{l,j}$ and $I_{l-1,j}$ are the feature maps of layers $l$ and $l-1$, respectively, $I_{l,j}^{i,x,y,z}$ denotes the value at position $(x,y,z)$ of the $j$th kernel in the $l$th layer, $W_{l,j,k}^{i,p,q,r}$ denotes the kernel value at location $(p,q,r)$ of the $k$th channel in the $j$th convolutional kernel connected to the $l$th convolution layer, $K_{l}$ denotes the number of 3-D feature cubes, $P_{l}$, $Q_{l}$, and $R_{l}$ are, respectively, the height, weight, and depth of the convolutional kernel, $(s_{1}, s_{2}, s_{3})$ denotes the size of stride in three dimensions and is set to 1, 1, and 1, and $\beta_{l,j}$ denotes the bias of the $j$th kernel in the $l$th layer.

The decoder of multiresolution CAE networks has the symmetric structure with respect to the encoder. The latent spatial-spectral features of each resolution are decoded to reconstruct each corresponding input HS cube by using the decoder networks. For each resolution decoder network, the latent spatial-spectral features $F_{ss}^{i}$ are fed into two deconvolution operations as

$$\bar{H} = f_{\text{deconv}2} \left[ f_{\text{deconv}1} \left( F_{ss}^{i} \right) \right], \quad i = 0, 1, 2 $$

where $\bar{H}$ is the reconstructed HS image of each resolution, and $f_{\text{deconv}1}$ and $f_{\text{deconv}2}$ are two deconvolution operations of the $i$th resolution.

A typical CAE has the common form of loss function, which minimizes the mean squared error between the input sample and the reconstructed output data. However, the mean squared error may generate oversmoothed effect at the edges. In this article, the $L_1$ loss function, instead of the mean squared error, is utilized to optimize each resolution 3-D CAE network. Besides, the spectral angle mapper (SAM) [37] further constrains the spectral error between the reference spectral vector of input HS data and the reconstructed one. The loss function of each resolution 3-D CAE network $(L_{1+SAM}^{3D})$ is designed as

$$\left( L_{1+SAM}^{3D} \right)^{i} = \left( L_{1}^{3D} \right)^{i} + \left( L_{SAM} \right)^{i} $$

where $(L_{1}^{3D})^{i}$ and $(L_{SAM})^{i}$ denote the $L_1$ loss and the SAM constraint of 3-D CAE for each resolution, respectively, $G$ denotes the number of training pairs, $\| \cdot \|_2$ denotes the $l_2$-norm, $H^{0} = H$, $H^{1} = \uparrow H$, and $H^{2} = \uparrow (\uparrow H)$ are the LR-HS image, the double-interpolated HS image, and the quadruple-interpolated HS image, respectively, $H_{g,a}$ denotes the spectral vector at the $a$th pixel of the HS image in the $g$th training sample, and $\alpha$ denotes a hyperparameter set as 0.001.

B. Multiresolution 2-D PAN Spatial Feature Extraction Network

The multiresolution 2-D spatial feature extraction networks are similar to the multiresolution 3-D spatial-spectral feature extraction networks. The multiresolution 3-D CAEs are applied to the HS image to extract the spatial-spectral features of the HS image, while the multiresolution 2-D CAEs are conducted in the spatial domain to extract the spatial features of the PAN image. With upsampling the HS image, the PAN image is downsampled twice to obtain two-time downsampled PAN image and four-time downsampled PAN image. When each scale PAN image is fed into the shallow 2-D CAEs, 32 filters of size $3 \times 3 \times 1$ are exploited to produce 32 feature maps, and 16 filters of size $3 \times 3 \times 32$ are then exploited to extract the hidden spatial feature maps. The activation functions used in the multiresolution 2-D CAEs are also ReLU. The produced spatial feature maps of each resolution can be expressed as

$$F_{ps}^{i} = \left\{ \begin{array}{ll}
    f_{\text{deconv}2} \left[ f_{\text{deconv}1} \left( P \right) \right], & i = 2 \\
    f_{\text{deconv}2} \left[ f_{\text{deconv}1} \left( \downarrow P \right) \right], & i = 1 \\
    f_{\text{deconv}2} \left[ f_{\text{deconv}1} \left( \downarrow \downarrow P \right) \right], & i = 0
\end{array} \right. $$

where $F_{ps}^{i}$ represents the spectral features of the PAN image at each scale, $\downarrow$ represents the downsampling operation, and $\downarrow \downarrow$ represents the double-downsampling operation.
\[ f_{2\text{conv}}^1 \text{ and } f_{2\text{conv}}^2 \text{ represent two 2-D convolution operations in the i\text{th} resolution PAN spatial feature extraction network. The decoders of multiresolution 2D CAEs also have the similar network architecture to that of the multiresolution 3D CAEs. The aggregated spatial features of each resolution are fed into two deconvolution operations to reconstruct each scale PAN image with ReLU being added between these two deconvolution layers for nonlinearity. In the multiresolution 2-D CAEs, we also use the L1 loss function to train the networks and add cross correlation (CC) that characterizes the spatial–spectral features.}

\[ \text{The loss function of each resolution 2-D CAEs} \quad L_i = \| \mu_i - \mu_i \|_2^2, \quad i = 0, 1, 2 \]

(6)

where \(L_i\) denotes the mean of the i\text{th} training sample. The first term in (6) measures the reconstruction error, while the second term ensures the spatial performance of the reconstructed PAN data.

C. Spatial–Spectral Feature-Based Multiresolution Residual HS Reconstruction Network

The multiresolution spatial–spectral features of three resolution HS images and the multiresolution spatial features of three resolution PAN images are completely extracted by using the 3-D CAE networks and 2-D CAE networks, with lower computational complexity. In order to effectively generate the pansharpened HR-HS image, as shown in Fig. 1, a multiresolution residual network is presented to reconstruct the HS image from the features extracted by six multiresolution CAE networks.

The extracted features of the HS and PAN images at the same resolution are aggregated to produce desired total spatial–spectral features at the corresponding resolution. For each resolution, the spatial–spectral features of HS image are concatenated with the spatial features of PAN image, which is given by

\[ F_i^{a} = C(F_i^{a_{1}}, F_i^{a_{2}}), \quad i = 0, 1, 2 \]

(7)

where \(F_i^a\) denotes the concatenated spatial–spectral features at each resolution, and \(C(\cdot)\) denotes the concatenation operation.

With the input as the concatenated spatial–spectral features at each scale \(F_i^{0}, F_i^{1}, \text{ and } F_i^{2}\), the proposed multiresolution residual reconstruction network aims at producing the HR pansharpened HS image \(H_F\). For each resolution, the spatial–spectral features at each resolution \(F_i\) are fed into one deconvolution layer and two residual convolution blocks in turn. In order to make full use of the extracted spatial–spectral features of each resolution and keep spatial–spectral information, the output features of the residual network at the lower resolution are upsampled by using the deconvolution layer and concatenated with the spatial–spectral features of the higher resolution as the input of the higher-resolution residual network. For each resolution, this process can be formulated as

\[ F_i^r = \begin{cases} F_i^r, & i = 0, 1 \\ H_F, & i = 2 \end{cases} \]

\[ F_i^{e} = \begin{cases} \text{f}_{\text{res}}^i \{ \text{f}_{\text{res}}^i \{ \text{f}_{\text{res}}^i (C(F_i^{a-1}, F_i^{r-1})) \} \}, & i = 0 \\ \text{f}_{\text{dec}}^i \{ \text{f}_{\text{res}}^i \{ \text{f}_{\text{res}}^i (C(F_i^{a-1}, F_i^{r-1})) \} \}, & i = 1 \end{cases} \]

\[ H_F = \text{f}_{\text{res}}^i \{ \text{f}_{\text{res}}^i \{ \text{f}_{\text{res}}^i (C(F_i^{a-1}, F_i^{r-1})) \} \}, \quad i = 2 \]

where \(F_i^r\) is the middle output features in the network, \(H_F\) is the HR pansharpened HS image, \(f_{\text{dec}}^i\) and \(f_{\text{res}}^i\) are the convolution and deconvolution operations of the \(i\)th resolution, and \(f_{\text{res}}^i\) and \(f_{\text{dec}}^i\) are two residual deconvolution blocks of the \(i\)th resolution. The L1-norm is utilized to train the multiresolution residual network with a loss function defined as

\[ L_{\text{resid}} = \frac{1}{E} \sum_{e=1}^{E} \| R_e - H_{F,e} \|_1 \]

(10)

where \(R \in M \times N \times c\) denotes the reference HS image, \(E\) denotes the number of training sample, and \(R_e\) and \(H_{F,e}\) are the \(e\)th sample of \(R\) and \(H_F\), respectively.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Settings

1) Data Sets: Three widely used data sets are applied for the performance evaluation of our proposed MSSL method. For network training, the input samples are obtained by three steps: 1) simulate the input HS and PAN images from the reference HS image according to Wald’s protocol [38]; 2) partition the simulated HS and PAN images into \(K_1\) nonoverlapped patches; and 3) enlarge the training samples by further extracting \(K_2\) patches from each nonoverlapped patch at a fixed stride with partial overlapping. It should be noted that there is no overlap between training and testing samples. A detailed introduction is given as follows.

Pavia Center: The Pavia Center data set was collected by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor ranging from the wavelengths of 430–860 nm, over Pavia, Italy. A subimage of size 960 × 640 × 102 is selected as a reference HS image in this article. Parameters \(K_1\) and \(K_2\) are set as 12 and 21, respectively.

Houston: The Houston data set was captured using an ITRES-CASI 1500 HS sensor over the University of Houston.
TABLE I
SIZE OF REFERENCE, SIMULATED, TRAINING, AND TESTING SAMPLES OF THREE DATA SETS. FOR HS, THE FOUR DIMENSIONS DENOTE PATCH NUMBER, HEIGHT, WIDTH, AND BAND, RESPECTIVELY. FOR PAN, THE THREE DIMENSIONS DENOTE PATCH NUMBER, HEIGHT, AND WIDTH, RESPECTIVELY. NOTE THAT THE PATCH NUMBER OF 1 IS OMITTED HERE.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Reference HS</th>
<th>Simulated HS</th>
<th>Simulated PAN</th>
<th>Training HS</th>
<th>Training PAN</th>
<th>Testing HS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pavia Center</td>
<td>960 × 640 × 102</td>
<td>240 × 160 × 102</td>
<td>960 × 640</td>
<td>6K&lt;sub&gt;2&lt;/sub&gt; × 40 × 40 × 102</td>
<td>6K&lt;sub&gt;2&lt;/sub&gt; × 160 × 160</td>
<td>3K&lt;sub&gt;2&lt;/sub&gt; × 40 × 40 × 102</td>
</tr>
<tr>
<td>Houston</td>
<td>320 × 1280 × 144</td>
<td>80 × 320 × 144</td>
<td>320 × 1280</td>
<td>6K&lt;sub&gt;2&lt;/sub&gt; × 40 × 40 × 144</td>
<td>6K&lt;sub&gt;2&lt;/sub&gt; × 160 × 160</td>
<td>2K&lt;sub&gt;2&lt;/sub&gt; × 40 × 40 × 144</td>
</tr>
<tr>
<td>Botswana</td>
<td>1280 × 240 × 145</td>
<td>320 × 60 × 145</td>
<td>1280 × 240</td>
<td>6K&lt;sub&gt;2&lt;/sub&gt; × 40 × 40 × 145</td>
<td>6K&lt;sub&gt;2&lt;/sub&gt; × 160 × 160</td>
<td>2K&lt;sub&gt;2&lt;/sub&gt; × 40 × 40 × 145</td>
</tr>
</tbody>
</table>

Fig. 2. Visual results of the Pavia Center data set. (a) Reference HS image. (b) GS. (c) GSA. (d) SFIM. (e) MGH. (f) GFPCA. (g) CNMF. (h) BSF. (i) DDLPS. (j) MSSL.

The campus and its neighboring urban area, Houston, USA. This data set originally contains 349 × 1905 pixels with a spatial resolution of 2.5 m and 144 bands covering from 380 to 1050 nm. For convenience, a subimage of size 320 × 1280 × 144 is cropped as reference HS in this experiment. K<sub>1</sub> and K<sub>2</sub> are set as 8 and 41, respectively.

Botswana: The Botswana data set was acquired by the Hyperion sensor loaded on Earth Observing One (EO-1) satellite over Okavango Delta, Botswana. These data have 1280 × 240 pixels with a spatial resolution of 30 m and 145 spectral bands ranging from 400 to 2500 nm after removing several low-SNR and noisy bands. K<sub>1</sub> and K<sub>2</sub> are also set as 8 and 41, respectively. Table I lists the size of training and testing sets of the three applied data sets in detail.

2) Evaluation Criterion: In this article, we evaluate the performance of the proposed method and other competitors from both qualitative and quantitative viewpoints. The visualization of the fused HS image and two kinds of error images (i.e., difference values and SAM images) between the fused HS and reference HS image is provided for subjective evaluation. On the other hand, four popular indexes, including CC [37], SAM, root mean squared error (RMSE), and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [39], are exploited for quantitative evaluation. The larger the CC value is, the better the spatial information injects. The smaller the RMSE and ERGAS are, the better the pansharpening method performs comprehensively.

3) Implementation: Since the multiresolution 3-D CAEs and 2-D CAEs are unsupervised neural networks, all samples are utilized to train the multiresolution 3-D CAEs and 2-D CAEs. The multiresolution residual HS reconstruction network is optimized using the training samples shown in Table I. All the experiments are implemented based on the PyTorch.

TABLE II
AVERAGE QUANTITATIVE ASSESSMENT AND COMPUTATIONAL COMPLEXITY FOR THE PAVIA CENTER DATA SET

<table>
<thead>
<tr>
<th>Method</th>
<th>CC</th>
<th>SAM</th>
<th>RMSE</th>
<th>ERGAS</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td>0.9314</td>
<td>6.5693</td>
<td>0.0362</td>
<td>5.5932</td>
<td>0.3090</td>
</tr>
<tr>
<td>GSA</td>
<td>0.9388</td>
<td>7.2563</td>
<td>0.0315</td>
<td>4.8912</td>
<td>0.3677</td>
</tr>
<tr>
<td>SFIM</td>
<td>0.9566</td>
<td>6.0115</td>
<td>0.0272</td>
<td>4.0610</td>
<td>0.3121</td>
</tr>
<tr>
<td>MGH</td>
<td>0.9606</td>
<td>5.9928</td>
<td>0.0260</td>
<td>3.7381</td>
<td>0.4211</td>
</tr>
<tr>
<td>GFPCA</td>
<td>0.8555</td>
<td>9.2956</td>
<td>0.0545</td>
<td>8.1331</td>
<td>1.0540</td>
</tr>
<tr>
<td>CNMF</td>
<td>0.9044</td>
<td>7.1717</td>
<td>0.0407</td>
<td>6.1798</td>
<td>1.7534</td>
</tr>
<tr>
<td>BSF</td>
<td>0.9141</td>
<td>8.7428</td>
<td>0.0396</td>
<td>6.1056</td>
<td>28.4442</td>
</tr>
<tr>
<td>DDLPS</td>
<td>0.9356</td>
<td>6.8955</td>
<td>0.0335</td>
<td>5.1023</td>
<td>0.3118</td>
</tr>
<tr>
<td>MSSL</td>
<td>0.9769</td>
<td>4.6019</td>
<td>0.0183</td>
<td>2.8741</td>
<td>0.1772</td>
</tr>
</tbody>
</table>
framework with NVIDIA GeForce GTX 2080Ti GPU. The epoch and batch size are set to 300 and 7, respectively. The learning rate of the multiresolution 3-D CAEs and 2-D CAEs is $10^{-3}$. The initial learning rate of the reconstruction network is set to $10^{-3}$, and the learning rate descent factor is 0.1 every 150 epochs.

B. Results and Analysis

In order to comprehensively evaluate the performance of the proposed MSSL method, several popular HS pansharpening methods are applied for comparison, including seven frequently used approaches, i.e., GS [13], GSA [14], SFIM [20], MGH [22], GFPCA [40], CNMF [27], and BSF [24], and one deep-learning-based approach called DDLPS [41]. The subjective and objective results on three test HS data sets are given in the following.

1) Pavia Center Data Set: Fig. 2 randomly visualizes one pansharpened HS patch obtained by different methods on the Pavia Center data set. It can be obviously observed that the reconstructed HS images obtained by GS, GSA, GFPCA, CNMF, BSF, and DDLPS have highly difference in color from the reference HS image, especially the hybrid GFPCA. Moreover, the visual results generated by GFPCA and BSF also look really blurry due to insufficient injection of spatial
Although two MRA-based methods, SFIM and MGH, perform better in spectral information preservation, there exist slight spatial distortions. In conclusion, the proposed MSSL method performs visually superior to other methods. As shown in Table II, the proposed MSSL method achieves the best results on all indexes. Moreover, MSSL outperforms the second MGH method in spatial enhancement and spectral preservation, mainly due to the integration of spatial information and spectral information and application of the SAM and CC loss.

2) Houston Data Set: Fig. 4 shows one reconstructed HS patch of the Houston data set obtained by all comparing methods. It is obvious that the pseudocolor images reconstructed by GS, GSA, GFPCA, CNMF, and BSF look relatively different from the reference image, which indicates that these methods cannot deal with the issue of spectral distortions well. In addition, some of them perform worse in spatial information injection, such as GFPCA and BSF. Compared with SFMI, MGH, and DDLPS, it is difficult to evaluate the pansharpening performance by sole visualization. Fig. 5 displays the error images (SAM) between the fused HS images and a reference image of all competitors. It can be clearly observed that our proposed MSSL has smaller values than SFIM, MGH, and DDLPS in most areas, especially the regions with complex spatial structures. This is mainly due to the high-level feature extraction at multiple resolutions. This fact demonstrates that MSSL outperforms other methods. Furthermore, quantitative evaluation results are provided in Table III. By analyzing these results, we can conclude that the proposed MSSL outperforms the other methods and obtains a significant spatial improvement.

3) Botswana Data Set: Fig. 6 displays the pseudocolor images of one HS patch obtained by all comparing methods on the Botswana data set. By observing from Fig. 6, we can see that, for this data set, most of the methods suffer from spectral distortions, such as GS, GSA, GFPCA, and BSF, since the color of the visual results remains different from that of the reference HS image. The visual result generated by CNMF is relatively blurry, indicating the loss of discriminative spatial information during reconstruction. Despite having a preeminent spatial quality, SFIM, MGH, and DDLPS show spectral distortions in some regions. From the visual analysis, MSSL effectively improves spatial performance while maintaining spectral fidelity. Fig. 7 shows the error images (difference of pixel values) of different methods on the Botswana data set. It is apparent that our proposed MSSL method can reconstruct the HS image with the least error. Table IV gives
the quantitative results for the Botswana data set. It can be seen that MSSL outperforms other methods from the viewpoints of spatial information injection and spectral distortion reduction, which completely complies with the subjective results in Fig. 6. In conclusion, the experimental results and analysis indicate that the proposed MSSL method can achieve superior pansharpening performance, and further analysis is given in the next section to verify the effectiveness of each applied component.

4) Computational Complexity: This section provides the computational complexity analysis of different competing methods on three HS data sets. All the experiments are tested on MATLAB (R2018b) with Intel Core i7-9750H CPU@2.60-GHz CPU and NVIDIA GeForce GTX 2080Ti GPU. Tables II–IV tabulate the computational costs of different competitors on three data sets. According to Tables II–IV, the proposed MSSL achieves high computation efficacy. The MSSL method ranks first place in terms of running time for the Pavia Center and Houston data sets. The MSSL method is faster than the GSA, MGH, GFPCA, and DDLPS algorithms for three data sets. Among all the competing methods, CNMF and BSF are the most time-consuming algorithms. Although the computational time of MSSL is slightly longer than the GS and SFIM methods for the Botswana data set, MSSL could obtain better performance at the expense of an acceptable computational cost. The proposed method designs several
shallow multiresolution networks to extract the spatial and spectral features of the HS and PAN images, which has a low computational cost. By comparing and analyzing the computational costs of each method, the fact that the proposed MSSL methods are computationally efficient confirms the advantage of low computational cost.

### C. Component Analysis

In this article, we verify the effectiveness of the proposed MSSL method from two aspects as follows.

1) Impacts of Loss and Multiresolution on Pansharpening:

In order to analyze the importance and effectiveness of loss function and multiresolution of MSSL, 12 cases are considered, as shown in Table V. In Table V, \( L_1^{3D} \) and \( L_1^{2D} \) denote the \( L_1 \) loss of 3-D CAE and 2-D CAE, respectively. \( L_1^{3D} + \text{SAM} \) and \( L_1^{2D} + \text{CC} \) represent the improved loss of 3-D CAE and 2-D CAE, respectively [see (4) and (6)]. \( s = 4 \), \( s = 1, 4 \), and \( s = 1, 2, 4 \) (see Fig. 1) are three model settings and represent that the proposed model is composed of single resolution \( (s = 4) \), double resolutions \( (s = 1, 4) \), and triple resolutions \( (s = 1, 2, 4) \), respectively.

Fig. 8 shows the performance of different model settings with different losses on the Pavia Center data set. According to Fig. 8, a significant improvement can be achieved from the proposed model with single resolution \( (s = 4) \) to the proposed model with double resolutions \( (s = 1, 4) \). It is because the model with the single resolution \( (s = 4) \) feeds the interpolated HS image and the PAN image into 3-D CAE and 2-D CAE and directly upsampling the HS image by four times may lose spectral information of the original LR-HS image. The model with double resolutions \( (s = 1, 4) \) supplements the lossless and sufficient spectral information provided by \( s = 1 \) and obtains a significant improvement. In comparison, the improvement from the proposed model with double resolutions \( (s = 1, 4) \) to the proposed model with triple resolutions \( (s = 1, 2, 4) \) is less significant.
TABLE V
ALL POSSIBLE COMBINATIONS BETWEEN LOSS FUNCTIONS AND RESOLUTIONS

<table>
<thead>
<tr>
<th>Loss in 3D CAE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1^{AD}$ ($L_1$ loss)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$L_1^{AD}$ at 1 and 4 (improved loss)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Loss in 2D CAE</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>$L_1^{2D}$ ($L_1$ loss)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$L_1^{2D}$ at 1 and 4 (improved loss)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Resolution</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>$s = 1$ (single resolution)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$s = 1, 4$ (double resolutions)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$s = 1, 2, 4$ (triple resolutions)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

is relatively unobvious. This is because the spatial–spectral features extracted from the middle resolution ($s = 2$) are supplemented by the complementary spatial–spectral features to the model with double resolutions ($s = 1$ and $s = 4$), which brings performance improvement. However, the spatial–spectral features extracted from $s = 2$ also contain redundant information with the spatial–spectral features extracted from $s = 1$ and $s = 4$, which leads to limited improvement. In other words, the model with double resolutions ($s = 1$ and $s = 4$) extracts most spatial–spectral features, and the model with triple resolutions ($s = 1, s = 2$ and $s = 4$) supplants the features extracted from $s = 2$ to provide a small amount of complementary spatial–spectral features. Overall, it can be observed from Fig. 8 that the fusion performance gradually improves as the resolution increases. As for the loss function, 2-D CAE and 3-D CAE with improved losses achieve better pansharpening performance than the other three combinations, which indicates the effectiveness of the improved losses. By contrast, the multiresolution networks with two improved loss functions obtain the best fusion performance.

2) Network Structure Analysis: Considering the 3-D cubic structure of HS images, this article applies 3-D CAE instead of 2-D CAE for spatial–spectral feature extraction of HS images. To verify its rationality, a comparison with 2-D CAE is performed. Moreover, we also evaluate the performance and computational complexity of our MSSL with different numbers of convolution/deconvolution layers in 3-D CAE (set as the same in 2-D CAE for spatial feature extraction). Fig. 9 provides the experimental results on the Pavia Center data set. Taking $2D_b = 2$ as an example, “$2 − D$” denotes that 2-D convolution/deconvolution is applied for spatial–spectral feature extraction of HS image, and the subscript “$b = 2$” is the number of convolution/deconvolution layers in 3-D CAE/2-D CAE. It should be noted that some indexes are scaled with a specific constant for clarity. By observing from Fig. 9, we can conclude two facts: 1) 2-D convolution/deconvolution operation is less suitable for HS image processing than 3-D and 2) more complex network does not mean better performance since the imbalance between increasing learnable parameters and limited training samples would generally lead to overfitting problem. $3D_{b=2}$ obtains acceptable computational cost compared with $3D_{b=1}$ and much better pansharpening performance than $3D_{b=1}$ and $3D_{b=3}$. Thus, 3-D CAE and 2-D CAE with two convolution and deconvolution layers are applied in the proposed MSSL method to achieve the best effectiveness–efficiency tradeoff.

IV. Conclusion

In this article, we propose an MSSL method for the fusion of HS and PAN images, which shows good generalization performance on different data sets. The MSSL designs several shallow multiresolution networks to extract adequate spatial–spectral features of the HS and PAN images. In order to reduce computation cost, the proposed MSSL upsamples the low spatial resolution HS image and designs shallow multiresolution 3-D CAEs that exploit spectral constraint to obtain spatial–spectral features of the HS image. Meanwhile, the MSSL downsamples the high spatial resolution PAN image and introduces shallow multiresolution 2-D CAEs with a spatial constraint to extract spatial features of the PAN image. Then, a multiresolution residual reconstruction network is proposed to reconstruct the fused HS image by learning the extracted spatial–spectral features from different resolutions. Extensive experiments validate the superiority of the proposed MSSL with evident performance improvements over competitors, both quantitatively and perceptually.

REFERENCES

Yanzi Shi (Student Member, IEEE) received the B.S. degree from the School of Telecommunications Engineering, Xidian University, Xi’an, China, in 2017, where she is pursuing the Ph.D. degree with the State Key Laboratory of Integrated Service Networks.

Her research interests include hyperspectral target detection, feature extraction, urban remote sensing, and deep learning.

Weiying Xie (Member, IEEE) received the B.S. degree in electronic information science and technology from the University of Jinan, Jinan, China, in 2011, the M.S. degree in communication and information systems from Lanzhou University, Lanzhou, China, in 2014, and the Ph.D. degree in communication and information systems from Xidian University, Xi’an, China, in 2017.

She is an Associate Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. She has published more than 30 articles in refereed journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, Neural Networks, and Pattern Recognition. Her research interests include neural networks, machine learning, hyperspectral image processing, and high-performance computing.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi’an, China, in 1999 and 2002, respectively.

He joined the School of Telecommunications Engineering, Xidian University, in 1999, where he is a Professor. He is also the Director of the Image Coding and Processing Center, State Key Laboratory of Integrated Service Networks, Xidian University. His research interests include image and video processing, hyperspectral image processing, and high-performance computing.

Xianyun Wu was born in 1985. He received the Ph.D. degree in telecommunication and information systems from Xidian University, Xi’an, China, in 2012.

He is an Associate Professor with Xidian University. His research interests include remote sensing, image compression and processing, and high-performance computing.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland, Baltimore County, Baltimore, MD, USA, in 2000.

She is a Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is also a fellow of the SPIE-International Society for Optics and Photonics. She was a recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society. She has served as the Co-Chair of the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society from 2009 to 2013 and the Chair of the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She is also the General Chair for the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing in Shanghai, China, in 2012. She has served as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the Journal of Applied Remote Sensing, and the IEEE SIGNAL PROCESSING LETTERS. Since 2016, she has been the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.